This seminar treats power and the various factors that affect power on both a conceptual and a mechanical level. While we will not cover the formulas needed to actually run a power analysis, later on we will discuss some of the software packages that can be used to conduct power analyses.

OK, let's start off with a basic definition of what a power is. Power is the probability of detecting an effect, given that the effect is really there. In other words, it is the probability of rejecting the null hypothesis when it is in fact false. For example, let's say that we have a simple study with drug A and a placebo group, and that the drug truly is effective; the power is the probability of finding a difference between the two groups. So, imagine that we had a power of .8 and that this simple study was conducted many times. Having power of .8 means that 80% of the time, we would get a statistically significant difference between the drug A and placebo groups. This also means that 20% of the times that we run this experiment, we will not obtain a statistically significant effect between the two groups, even though there really is an effect in reality.

There are several of reasons why one might do a power analysis. Perhaps the most common use is to determine the necessary number of subjects needed to detect an effect of a given size. Note that trying to find the absolute, bare minimum number of subjects needed in the study is often not a good idea. Additionally, power analysis can be used to determine power, given an effect size and the number of subjects available. You might do this when you know, for example, that only 75 subjects are available (or that you only have the budget for 75 subjects), and you want to know if you will have enough power to justify actually doing the study. In most cases, there is really no point to conducting a study that is seriously underpowered. Besides the issue of the number of necessary subjects, there are other good reasons for doing a power analysis. For example, a power analysis is often required as part of a grant proposal. And finally, doing a power analysis is often just part of doing good research. A power analysis is a good way of making sure that you have thought through every aspect of the study and the statistical analysis before you start collecting data.

Despite these advantages of power analyses, there are some limitations. One limitation is that power analyses do not typically generalize very well. If you change the methodology used to collect the data or change the statistical procedure used to analyze the data, you will most likely have to redo the power analysis. In some cases, a power analysis might suggest a number of subjects that is inadequate for the statistical procedure. For example, a power analysis might suggest that you need 30 subjects for your logistic regression, but logistic regression, like all maximum likelihood procedures, require much larger sample sizes. Perhaps the most important limitation is that a standard power analysis gives you a "best case scenario" estimate of the necessary number of subjects needed to detect the effect. In most cases, this "best case scenario" is based on assumptions and educated guesses. If any of these assumptions or

guesses are incorrect, you may have less power than you need to detect the effect. Finally, because power analyses are based on assumptions and educated guesses, you often get a range of the number of subjects needed, not a precise number. For example, if you do not know what the standard deviation of your outcome measure will be, you guess at this value, run the power analysis and get X number of subjects. Then you guess a slightly larger value, rerun the power analysis and get a slightly larger number of necessary subjects. You repeat this process over the plausible range of values of the standard deviation, which gives you a range of the number of subjects that you will need.

After all of this discussion of power analyses and the necessary number of subjects, we need to stress that power is not the only consideration when determining the necessary sample size. For example, different researchers might have different reasons for conducting a regression analysis. One might want to see if the regression coefficient is different from zero, while the other wants to get a very precise estimate of the regression coefficient with a very small confidence interval around it. This second purpose requires a larger sample size than does merely seeing if the regression coefficient is different from zero. Another consideration when determining the necessary sample size is the assumptions of the statistical procedure that is going to be used. The number of statistical tests that you intend to conduct will also influence your necessary sample size: the more tests that you want to run, the more subjects that you will need. You will also want to consider the representativeness of the sample, which, of course, influences the generalizability of the results. Unless you have a really sophisticated sampling plan, the greater the desired generalizability, the larger the necessary sample size. Finally, please note that most of what is in this presentation does not readily apply to people who are developing a sampling plan for a survey or psychometric analyses.

Definitions

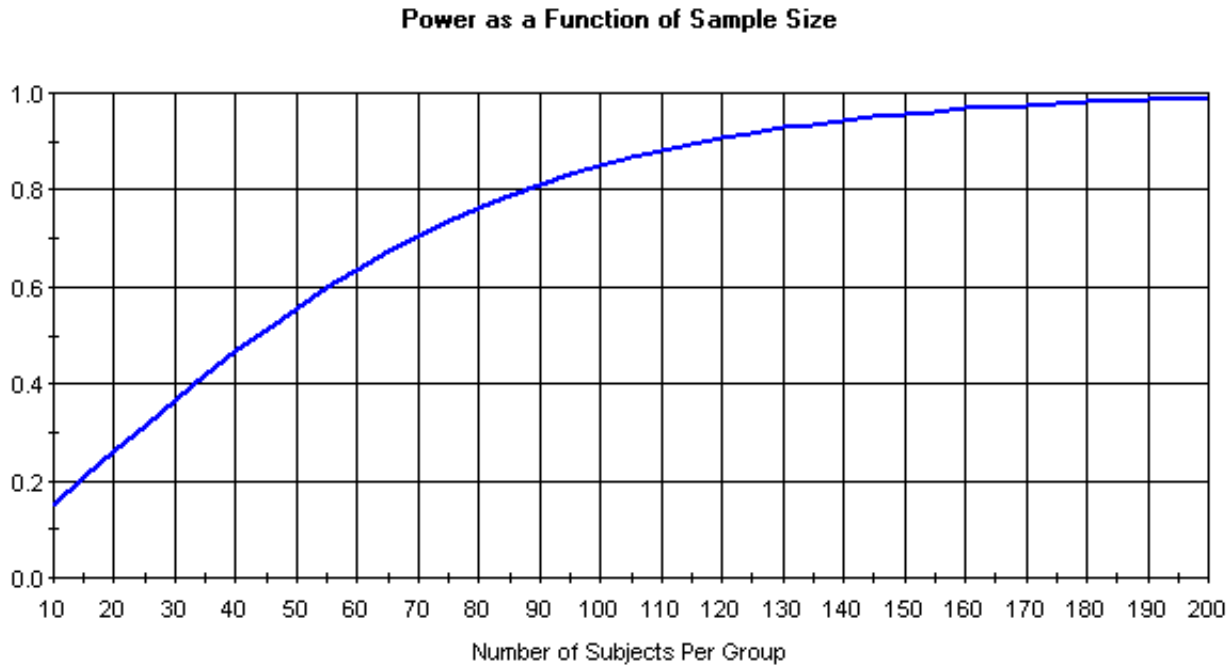
Before we move on, let's make sure we are all using the same definitions. We have already defined power as the probability of detecting a "true" effect, when the effect exists. Most recommendations for power fall between .8 and .9. We have also been using the term "effect size", and while intuitively it is an easy concept, there are lots of definitions and lots of formulas for calculating effect sizes. For example, the current APA manual has a list of more than 15 effect sizes, and there are more than a few books mostly dedicated to the calculation of effect sizes in various situations. For now, let's stick with one of the simplest definitions, which is that an effect size is the difference of two group means divided by the pooled standard deviation. Going back to our previous example, suppose the mean of the outcome variable for the drug A group was 10 and it was 5 for the placebo group. If the pooled standard deviation was 2.5, we would have and effect size which is equal to (10-5)/2.5 = 2 (which is a large effect size).

We also need to think about “statistically significance” versus “clinically relevant”. This issue comes up often when considering effect sizes. For example, for a given number of subjects, you might only need a small effect size to have a power of .9. But that effect size might correspond to a difference between the drug and placebo groups that isn’t clinically meaningful, say reducing blood pressure by two points. So even though you would have enough power, it still might not be worth doing the study, because the results would not be useful for clinicians.

There are a few other definitions that we will need later in this seminar. A Type I error occurs when the null hypothesis is true (in other words, there really is no effect), but you reject the null hypothesis. A Type II error occurs when the alternative hypothesis is correct, but you fail to reject the null hypothesis (in other words, there really is an effect, but you failed to detect it). Alpha inflation refers to the increase in the nominal alpha level when the number of statistical tests conducted on a given data set is increased.

When discussing statistical power, we have four inter-related concepts: power, effect size, sample size and alpha. These four things are related such that each is a function of the other three. In other words, if three of these values are fixed, the fourth is completely determined (Cohen, 1988, page 14). We mention this because, by increasing one, you can decrease (or increase) another. For example, if you can increase your effect size, you will need fewer subjects, given the same power and alpha level. Specifically, increasing the effect size, the sample size and/or alpha will increase your power.

While we are thinking about these related concepts and the effect of increasing things, let’s take a quick look at a standard power graph. (This graph was made in SPSS Sample Power, and for this example, we’ve used .61 and 4 for our two proportion positive values.)



We like these kinds of graphs because they make clear the diminishing returns you get for adding more and more subjects. For example, let's say that we have only 10 subjects per group. We can see that we have a power of about .15, which is really, really low. We add 50 subjects per group, now we have a power of about .6, an increase of .45. However, if we started with 100 subjects per group (power of about .8) and added 50 per group, we would have a power of .95, an increase of only .15. So each additional subject gives you less additional power. This curve also illustrates the "cost" of increasing your desired power from .8 to .9.

Knowing your research project

As we mentioned before, one of the big benefits of doing a power analysis is making sure that you have thought through every detail of your research project.

Now most researchers have thought through most, if not all, of the substantive issues involved in their research. While this is absolutely necessary, it often is not sufficient. Researchers also need to carefully consider all aspects of the experimental design, the variables involved, and the statistical analysis technique that will be used. As you will see in the next sections of this presentation, a power analysis is the union of substantive knowledge (i.e., knowledge about the subject matter), experimental or quasi-experimental design issues, and statistical analysis. Almost every aspect of the experimental design can affect power. For example, the type of control group that is used or the number of time points that are collected will affect how much power you have. So knowing about these issues and carefully considering your options is important. There are plenty of excellent books that cover these issues in detail, including

Shadish, Cook and Campbell (2002); Cook and Campbell (1979); Campbell and Stanley (1963); Brickman (2000a, 2000b); Campbell and Russo (2001); Webb, Campbell, Schwartz and Sechrest (2000); and Anderson (2001).

Also, you want to know as much as possible about the statistical technique that you are going to use. If you learn that you need to use a binary logistic regression because your outcome variable is 0/1, don't stop there; rather, get a sample data set (there are plenty of sample data sets on our web site) and try it out. You may discover that the statistical package that you use doesn't do the type of analysis that need to do. For example, if you are an SPSS user and you need to do a weighted multilevel logistic regression, you will quickly discover that SPSS doesn't do that (as of version 25), and you will have to find (and probably learn) another statistical package that will do that analysis. Maybe you want to learn another statistical package, or maybe that is beyond what you want to do for this project. If you are writing a grant proposal, maybe you will want to include funds for purchasing the new software. You will also want to learn what the assumptions are and what the "quirks" are with this particular type of analysis. Remember that the number of necessary subjects given to you by a power analysis assumes that all of the assumptions of the analysis have been met, so knowing what those assumptions are is important deciding if they are likely to be met or not.

The point of this section is to make clear that knowing your research project involves many things, and you may find that you need to do some research about experimental design or statistical techniques before you do your power analysis.

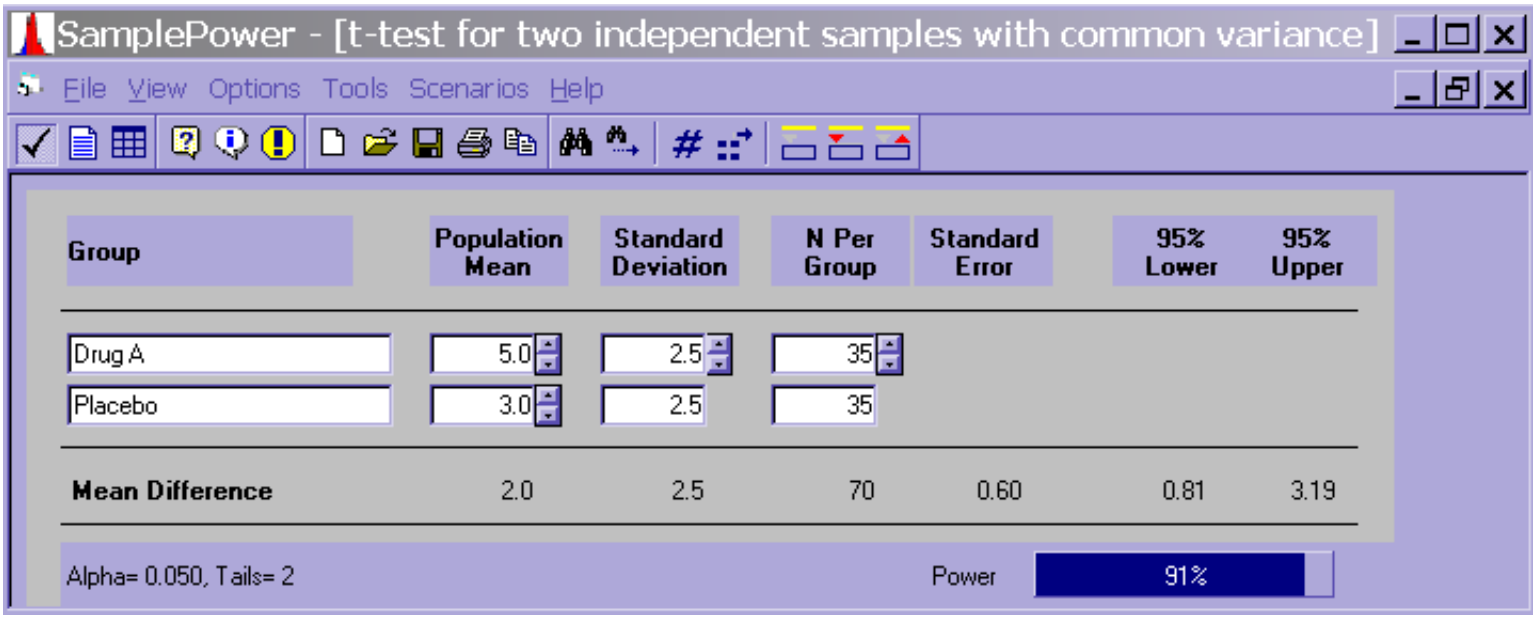
We want to emphasize that this is time and effort well spent. We also want to remind you that for almost all researchers, this is a normal part of doing good research. UCLA researchers are welcome and encouraged to come by walk-in consulting (/schedule/statistical-consulting-schedule/) at this stage of the research process to discuss issues and ideas, check out books and try out software.

What you need to know to do a power analysis

In the previous section, we discussed in general terms what you need to know to do a power analysis. In this section we will discuss some of the actual quantities that you need to know to do a power analysis for some simple statistics. Although we understand very few researchers test their main hypothesis with a t-test or a chi-square test, our point here is only to give you a flavor of the types of things that you will need to know (or guess at) in order to be ready for a power analysis.

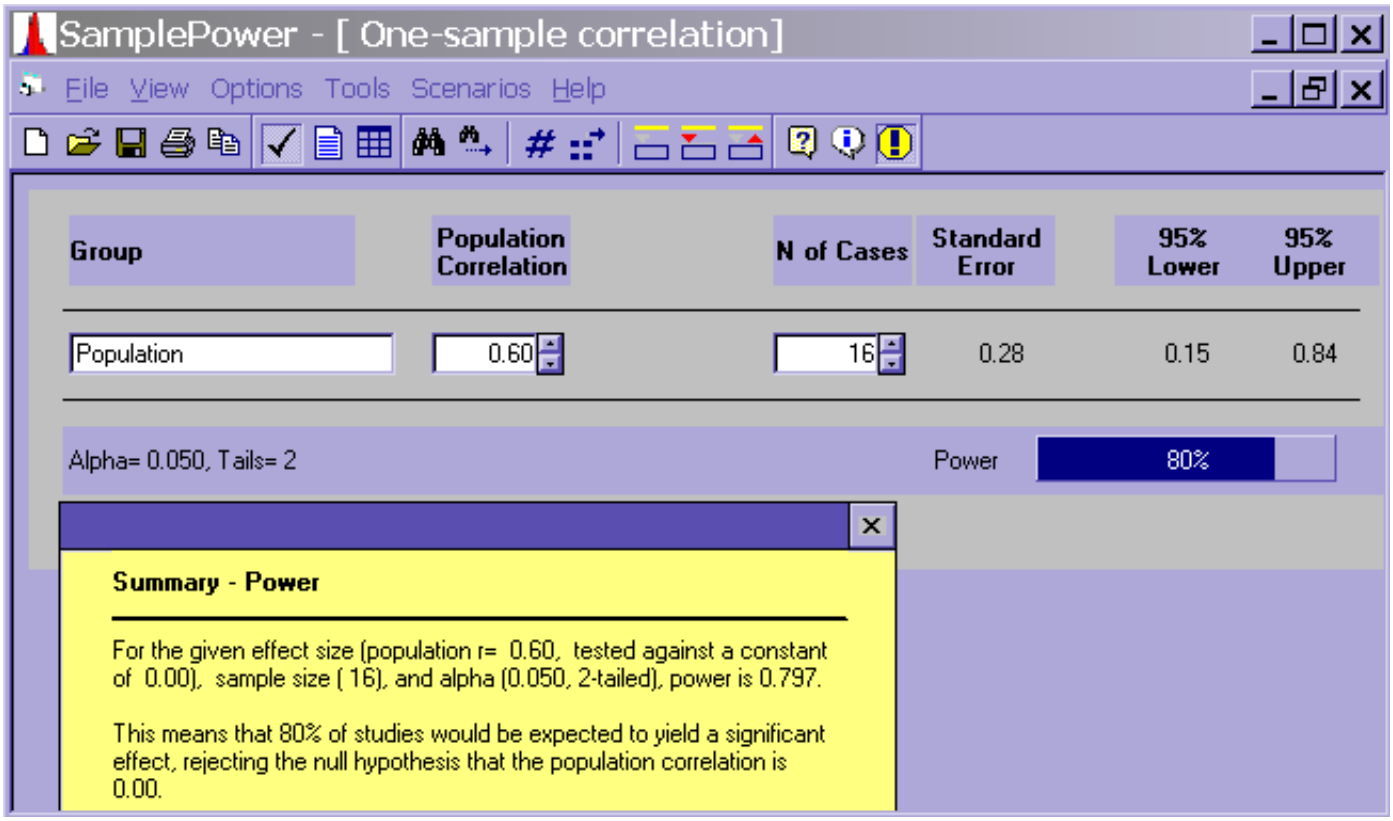
– For an independent samples t-test, you will need to know the population means of the two groups (or the difference between the means), and the population standard deviations of the two groups. So, using our example of drug A and placebo, we would need to know the difference in the means of the two groups, as well as the standard deviation for each group (because the group means and standard deviations are the best estimate that we have of those population values). Clearly, if we knew all of this, we wouldn't need to conduct the study. In reality, researchers make educated guesses at these values. We always recommend that you use several different values, such as decreasing the difference in the means and increasing the standard deviations, so that you get a range of values for the number of necessary subjects.

In SPSS Sample Power, we would have a screen that looks like the one below, and we would fill in the necessary values. As we can see, we would need a total of 70 subjects (35 per group) to have a power of .91 if we had a mean of 5 and a standard deviation of 2.5 in the drug A group, and a mean of 3 and a standard deviation of 2.5 in the placebo group. If we decreased the difference in the means and increased the standard deviations such that for the drug A group, we had a mean of 4.5 and a standard deviation of 3, and for the placebo group a mean of 3.5 and a standard deviation of 3, we would need 190 subjects per group, or a total of 380 subjects, to have a power of .90. In other words, seemingly small differences in means and standard deviations can have a huge effect on the number of subjects required.

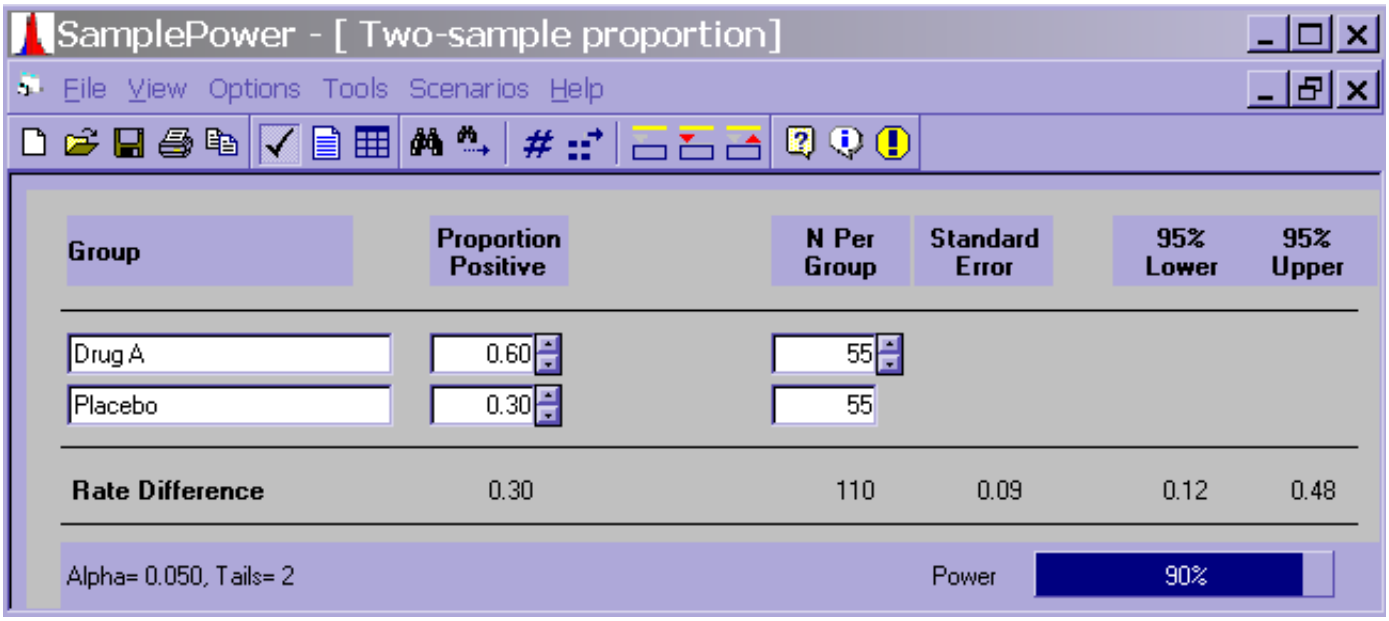


– For a correlation, you need to know/guess at the correlation in the population. This is a good time to remember back to an early stats class where they emphasized that correlation is a large *N* procedure (Chen and Popovich, 2002). If you guess that the population correlation is .6, a power analysis would suggest (with an alpha of .05 and for a power of .8) that you would need only 16 subjects. There are several points to be made here. First, common sense suggests that

N = 16 is pretty low. Second, a population correlation of .6 is pretty high, especially in the social sciences. Third, the power analysis assumes that all of the assumptions of the correlation have been met. For example, we are assuming that there is no restriction of range issue, which is common with Likert scales; the sample data for both variables are normally distributed; the relationship between the two variables is linear; and there are no serious outliers. Also, whereas you might be able to say that the sample correlation does not equal zero, you likely will not have a very precise estimate of the population correlation coefficient.



– For a chi-square test, you will need to know the proportion positive for both populations (i.e., rows and columns). Let's assume that we will have a 2 x 2 chi-square, and let's think of both variables as 0/1. Let's say that we wanted to know if there was a relationship between drug group (drug A/placebo) and improved health. In SPSS Sample Power, you would see a screen like this.



In order to get the .60 and the .30, we would need to know (or guess at) the number of people whose health improved in both the drug A and placebo groups.

We would also need to know (or guess at) either the number of people whose health did not improve in those two groups, or the total number of people in each group.

	Improved health (positive)	Not improved health	Row total
Drug A (positive)	33 (33/55 = .6)	22	55
Placebo	17 (17/55 = .3)	38	55
Column total	50	60	Grand Total = 110

– For an ordinary least squares regression, you would need to know things like the R^2 for the full and reduced model. For a simple logistic regression analysis with only one continuous predictor variable, you would need to know the probability of a positive outcome (i.e., the probability that the outcome equals 1) at the mean of the predictor variable and the probability of a positive outcome at one standard deviation above the mean of the predictor variable. Especially for the various types of logistic models (e.g., binary, ordinal and multinomial), you will need to think very carefully about your sample size, and information from a power analysis will only be part of your

considerations. For example, according to Long (1997, pages 53-54), 100 is a minimum sample size for logistic regression, and you want *at least* 10 observations per predictor. This does not mean that if you have only one predictor you need only 10 observations.

Also, if you have categorical predictors, you may need to have more observations to avoid computational difficulties caused by empty cells or cells with few observations. More observations are needed when the outcome variable is very lopsided; in other words, when there are very few 1s and lots of 0s, or vice versa. These cautions emphasize the need to know your data set well, so that you know if your outcome variable is lopsided or if you are likely to have a problem with empty cells.

The point of this section is to give you a sense of the level of detail about your variables that you need to be able to estimate in order to do a power analysis. Also, when doing power analyses for regression models, power programs will start to ask for values that most researchers are not accustomed to providing. Guessing at the mean and standard deviation of your response variable is one thing, but increments to R^2 is a metric in which few researchers are used to thinking. In our next section we will discuss how you can guestimate these numbers.

Obtaining the necessary numbers to do a power analysis

There are at least three ways to guestimate the values that are needed to do a power analysis: a literature review, a pilot study and using Cohen's recommendations. We will review the pros and cons of each of these methods. For this discussion, we will focus on finding the effect size, as that is often the most difficult number to obtain and often has the strongest impact on power.

Literature review: Sometimes you can find one or more published studies that are similar enough to yours that you can get a idea of the effect size. If you can find several such studies, you might be able to use meta-analysis techniques to get a robust estimate of the effect size. However, oftentimes there are no studies similar enough to your study to get a good estimate of the effect size. Even if you can find such an study, the necessary effect sizes or other values are often not clearly stated in the article and need to be calculated (if they can) based on the information provided.

Pilot studies: There are lots of good reasons to do a pilot study prior to conducting the actual study. From a power analysis prospective, a pilot study can give you a rough estimate of the effect size, as well as a rough estimate of the variability in your measures. You can also get some idea about where missing data might occur, and as we will discuss later, how you handle missing data can greatly affect your power. Other benefits of a pilot study include allowing you to identify coding problems, setting up the data base, and inputting the data for a practice analysis. This will allow you to determine if the data are input in the correct shape, etc.

Of course, there are some limitations to the information that you can get from a pilot study. (Many of these limitations apply to small samples in general.) First of all, when estimating effect sizes based on nonsignificant results, the effect size estimate will necessarily have an increased error; in other words, the standard error of the effect size estimate will be larger than when the result is significant. The effect size estimate that you obtain may be unduly influenced by some peculiarity of the small sample. Also, you often cannot get a good idea of the degree of missingness and attrition that will be seen in the real study. Despite these limitations, we strongly encourage researchers to conduct a pilot study. The opportunity to identify and correct "bugs" before collecting the real data is often invaluable. Also, because of the number of values that need to be guestimated in a power analysis, the precision of any one of these values is not that important. If you can estimate the effect size to within 10% or 20% of the true value, that is probably sufficient for you to conduct a meaningful power analysis, and such fluctuations can be taken into account during the power analysis.

Cohen's recommendations: Jacob Cohen has many well-known publications regarding issues of power and power analyses, including some recommendations about effect sizes that you can use when doing your power analysis. Many researchers (including Cohen) consider the use of such recommendations as a last resort, when a thorough literature review has failed to reveal any useful numbers and a pilot study is either not possible or not feasible. From Cohen (1988, pages 24-27):

– Small effect: 1% of the variance; $d = 0.25$ (too small to detect other than statistically; lower limit of what is clinically relevant)

– Medium effect: 6% of the variance; $d = 0.5$ (apparent with careful observation)

– Large effect: at least 15% of the variance; $d = 0.8$ (apparent with a superficial glance; unlikely to be the focus of research because it is too obvious)

Lipsey and Wilson (1993) did a meta analysis of 302 meta analyses of over 10,000 studies and found that the average effect size was .5, adding support to Cohen's recommendation that, as a last resort, guess that the effect size is .5 (cited in Bausell and Li, 2002). Sedlmeier and Gigerenzer (1989) found that the average effect size for articles in The Journal of Abnormal Psychology was a medium effect. According to Keppel and Wickens (2004), when you really have no idea what the effect size is, go with the smallest effect size of practical value. In other words, you need to know how small of a difference is meaningful to you. Keep in mind that research suggests that most researchers are overly optimistic about the effect sizes in their

research, and that most research studies are under powered (Keppel and Wickens, 2004; Tversky and Kahneman, 1971). This is part of the reason why we stress that a power analysis gives you a lower limit to the number of necessary subjects.

Factors that affect power

From the preceding discussion, you might be starting to think that the number of subjects and the effect size are the most important factors, or even the only factors, that affect power. Although effect size is often the largest contributor to power, saying it is the only important issue is far from the truth. There are at least a dozen other factors that can influence the power of a study, and many of these factors should be considered not only from the perspective of doing a power analysis, but also as part of doing good research. The first couple of factors that we will discuss are more "mechanical" ways of increasing power (e.g., alpha level, sample size and effect size). After that, the discussion will turn to more methodological issues that affect power.

1. Alpha level: One obvious way to increase your power is to increase your alpha (from .05 to say, .1). Whereas this might be an advisable strategy when doing a pilot study, increasing your alpha usually is not a viable option. We should point out here that many researchers are starting to prefer to use .01 as an alpha level instead of .05 as a crude attempt to assure results are clinically relevant; this alpha reduction reduces power.

1a. One- versus two-tailed tests: In some cases, you can test your hypothesis with a one-tailed test. For example, if your hypothesis was that drug A is better than the placebo, then you could use a one-tailed test. However, you would fail to detect a difference, even if it was a large difference, if the placebo was better than drug A. The advantage of one-tailed tests is that they put all of your power "on one side" to test your hypothesis. The disadvantage is that you cannot detect differences that are in the opposite direction of your hypothesis. Moreover, many grant and journal reviewers frown on the use of one-tailed tests, believing it is a way to feign significance (Stratton and Neil, 2004).

2. Sample size: A second obvious way to increase power is simply collect data on more subjects. In some situations, though, the subjects are difficult to get or extremely costly to run. For example, you may have access to only 20 autistic children or only have enough funding to interview 30 cancer survivors. If possible, you might try increasing the number of subjects in groups that do not have these restrictions, for example, if you are comparing to a group of normal controls. While it is true that, in general, it is often desirable to have roughly the same number of subjects in each group, this is not absolutely necessary. However, you get diminishing returns for additional subjects in the control group: adding an extra 100 subjects to the control group might not be much more helpful than adding 10 extra subjects to the control group.

3. Effect size: Another obvious way to increase your power is to increase the effect size. Of course, this is often easier said than done. A common way of increasing the effect size is to increase the experimental manipulation. Going back to our example of drug A and placebo, increasing the experimental manipulation might mean increasing the dose of the drug. While this might be a realistic option more often than increasing your alpha level, there are still plenty of times when you cannot do this. Perhaps the human subjects committee will not allow it, it does not make sense clinically, or it doesn't allow you to generalize your results the way you want to. Many of the other issues discussed below indirectly increase effect size by providing a stronger research design or a more powerful statistical analysis.

4. Experimental task: Well, maybe you can not increase the experimental manipulation, but perhaps you can change the experimental task, if there is one. If a variety of tasks have been used in your research area, consider which of these tasks provides the most power (compared to other important issues, such as relevancy, participant discomfort, and the like). However, if various tasks have not been reviewed in your field, designing a more sensitive task might be beyond the scope of your research project.

5. Response variable: How you measure your response variable(s) is just as important as what task you have the subject perform. When thinking about power, you want to use a measure that is as high in sensitivity and low in measurement error as is possible. Researchers in the social sciences often have a variety of measures from which they can choose, while researchers in other fields may not. For example, there are numerous established measures of anxiety, IQ, attitudes, etc. Even if there are not established measures, you still have some choice. Do you want to use a Likert scale, and if so, how many points should it have? Modifications to procedures can also help reduce measurement error. For example, you want to make sure that each subject knows exactly what he or she is supposed to be rating. Oral instructions need to be clear, and items on questionnaires need to be unambiguous to all respondents. When possible, use direct instead of indirect measures. For example, asking people what tax bracket they are in is a more direct way of determining their annual income than asking them about the square footage of their house. Again, this point may be more applicable to those in the social sciences than those in other areas of research. We should also note that minimizing the measurement error in your predictor variables will also help increase your power.

Just as an aside, most texts on experimental design strongly suggest collecting more than one measure of the response in which you are interested. While this is very good methodologically and provides marked benefits for certain analyses and missing data, it does complicate the power analysis.

6. Experimental design: Another thing to consider is that some types of experimental designs are more powerful than others. For example, repeated measures designs are virtually always more powerful than designs in which you only get measurements at one time. If you are already using a repeated measures design, increasing the number of time points a response variable is collected to at least four or five will also provide increased power over fewer data collections. There is a point of diminishing return when a researcher collects too many time points, though this depends on many factors such as the response variable, statistical design, age of participants, etc.

7. Groups: Another point to consider is the number and types of groups that you are using. Reducing the number of experimental conditions will reduce the number of subjects that is needed, or you can keep the same number of subjects and just have more per group. When thinking about which groups to exclude from the design, you might want to leave out those in the middle and keep the groups with the more extreme manipulations. Going back to our drug A example, let's say that we were originally thinking about having a total of four groups: the first group will be our placebo group, the second group would get a small dose of drug A, the third group a medium dose, and the fourth group a large dose. Clearly, much more power is needed to detect an effect between the medium and large dose groups than to detect an effect between the large dose group and the placebo group. If we found that we were unable to increase the power enough such that we were likely to find an effect between small and medium dose groups or between the medium and the large dose groups, then it would probably make more sense to run the study without these groups. In some cases, you may even be able to change your comparison group to something more extreme. For example, we once had a client who was designing a study to compare people with clinical levels of anxiety to a group that had subclinical levels of anxiety. However, while doing the power analysis and realizing how many subjects she would need to detect the effect, she found that she needed far fewer subjects if she compared the group with the clinical levels of anxiety to a group of "normal" people (a number of subjects she could reasonably obtain).

8. Statistical procedure: Changing the type of statistical analysis may also help increase power, especially when some of the assumptions of the test are violated. For example, as Maxwell and Delaney (2004) noted, "Even when ANOVA is robust, it may not provide the most powerful test available when its assumptions have been violated." In particular, violations of assumptions regarding independence, normality and heterogeneity can reduce power. In such cases, nonparametric alternatives may be more powerful.

9. Statistical model: You can also modify the statistical model. For example, interactions often require more power than main effects. Hence, you might find that you have reasonable power for a main effects model, but not enough power when the model includes interactions. Many

(perhaps most?) power analysis programs do not have an option to include interaction terms when describing the proposed analysis, so you need to keep this in mind when using these programs to help you determine how many subjects will be needed. When thinking about the statistical model, you might want to consider using covariates or blocking variables. Ideally, both covariates and blocking variables reduce the variability in the response variable. However, it can be challenging to find such variables. Moreover, your statistical model should use as many of the response variable time points as possible when examining longitudinal data. Using a change-score analysis when one has collected five time points makes little sense and ignores the added power from these additional time points. The more the statistical model "knows" about how a person changes over time, the more variance that can be pulled out of the error term and ascribed to an effect.

9a. Correlation between time points: Understanding the expected correlation between a response variable measured at one time in your study with the same response variable measured at another time can provide important and power-saving information. As noted previously, when the statistical model has a certain amount of information regarding the manner by which people change over time, it can enhance the effect size estimate. This is largely dependent on the correlation of the response measure over time. For example, in a before-after data collection scenario, response variables with a .00 correlation from before the treatment to after the treatment would provide no extra benefit to the statistical model, as we can't better understand a subject's score by knowing how he or she changes over time. Rarely, however, do variables have a .00 correlation on the same outcomes measured at different times. It is important to know that outcome variables with larger correlations over time provide enhanced power when used in a complimentary statistical model.

10. Modify response variable: Besides modifying your statistical model, you might also try modifying your response variable. Possible benefits of this strategy include reducing extreme scores and/or meeting the assumptions of the statistical procedure. For example, some response variables might need to be log transformed. However, you need to be careful here. Transforming variables often makes the results more difficult to interpret, because now you are working in, say, a logarithm metric instead of the metric in which the variable was originally measured. Moreover, if you use a transformation that adjusts the model too much, you can loose more power than is necessary. Categorizing continuous response variables (sometimes used as a way of handling extreme scores) can also be problematic, because logistic or ordinal logistic regression often requires many more subjects than does OLS regression. It makes sense that categorizing a response variable will lead to a loss of power, as information is being "thrown away."

11. Purpose of the study: Different researchers have different reasons for conducting research. Some are trying to determine if a coefficient (such as a regression coefficient) is different from zero. Others are trying to get a precise estimate of a coefficient. Still others are replicating research that has already been done. The purpose of the research can affect the necessary sample size. Going back to our drug A and placebo study, let's suppose our purpose is to test the difference in means to see if it equals zero. In this case, we need a relatively small sample size. If our purpose is to get a precise estimate of the means (i.e., minimizing the standard errors), then we will need a larger sample size. If our purpose is to replicate previous research, then again we will need a relatively large sample size. Tversky and Kahneman (1971) pointed out that we often need more subjects in a replication study than were in the original study. They also noted that researchers are often too optimistic about how much power they really have. They claim that researchers too readily assign "causal" reasons to explain differences between studies, instead of sampling error. They also mentioned that researchers tend to underestimate the impact of sampling and think that results will replicate more often than is the case.

12. Missing data: A final point that we would like to make here regards missing data. Almost all researchers have issues with missing data. When designing your study and selecting your measures, you want to do everything possible to minimize missing data. Handling missing data via imputation methods can be very tricky and very time-consuming. If the data set is small, the situation can be even more difficult. In general, missing data reduces power; poor imputation methods can greatly reduce power. If you have to impute, you want to have as few missing data points on as few variables as possible. When designing the study, you might want to collect data specifically for use in an imputation model (which usually involves a different set of variables than the model used to test your hypothesis). It is also important to note that the default technique for handling missing data by virtually every statistical program is to remove the entire case from an analysis (i.e., listwise deletion). This process is undertaken even if the analysis involves 20 variables and a subject is missing only one datum of the 20. Listwise deletion is one of the biggest contributors to loss of power, both because of the omnipresence of missing data and because of the omnipresence of this default setting in statistical programs (Graham et al., 2003).

This ends the section on the various factors that can influence power. We know that was a lot, and we understand that much of this can be frustrating because there is very little that is "black and white". We hope that this section made clear the close relationship between the experimental design, the statistical analysis and power.

Cautions about small sample sizes and sampling variation

We want to take a moment here to mention some issues that frequently arise when using small samples. (We aren't going to put a lower limit on what we mean be "small sample size.") While there are situations in which a researcher can either only get or afford a small number of subjects, in most cases, the researcher has some choice in how many subjects to include. Considerations of time and effort argue for running as few subjects as possible, but there are some difficulties associated with small sample sizes, and these may outweigh any gains from the saving of time, effort or both. One obvious problem with small sample sizes is that they have low power. This means that you need to have a large effect size to detect anything. You will also have fewer options with respect to appropriate statistical procedures, as many common procedures, such as correlations, logistic regression and multilevel modeling, are not appropriate with small sample sizes. It may also be more difficult to evaluate the assumptions of the statistical procedure that is used (especially assumptions like normality). In most cases, the statistical model must be smaller when the data set is small. Interaction terms, which often test interesting hypotheses, are frequently the first casualties. Generalizability of the results may also be comprised, and it can be difficult to argue that a small sample is representative of a large and varied population. Missing data are also more problematic; there are a reduced number of imputations methods available to you, and these are not considered to be desirable imputation methods (such as mean imputation). Finally, with a small sample size, alpha inflation issues can be more difficult to address, and you are more likely to run as many tests as you have subjects.

While the issue of sampling variability is relevant to all research, it is especially relevant to studies with small sample sizes. To quote Murphy and Myors (2004, page 59), "The lack of attention to power analysis (and the deplorable habit of placing too much weight on the results of small sample studies) are well documented in the literature, and there is no good excuse to ignore power in designing studies." In an early article entitled *The Law of Small Numbers*, Tversky and Kahneman (1971) stated that many researchers act like the Law of Large Numbers applies to small numbers. People often believe that small samples are more representative of the population than they really are.

The last two points to be made here is that there is usually no point to conducting an underpowered study, and that underpowered studies can cause chaos in the literature because studies that are similar methodologically may report conflicting results.

Software

We will briefly discuss some of the programs that you can use to assist you with your power analysis. Most programs are fairly easy to use, but you still need to know effect sizes, means, standard deviations, etc.

Among the programs specifically designed for power analysis, we use SPSS Sample Power, PASS and GPower. These programs have a friendly point-and-click interface and will do power analyses for things like correlations, OLS regression and logistic regression. We have also started using Optimal Design for repeated measures, longitudinal and multilevel designs. We should note that Sample Power is a stand-alone program that is sold by SPSS; it is not part of SPSS Base or an add-on module. PASS can be purchased directly from NCSS at http://www.ncss.com/index.htm (http://www.ncss.com/index.htm) . GPower (please see GPower (http://www.gpower.hhu.de/en.html) for details) and Optimal Design (please see http://sitemaker.umich.edu/group-based/home (http://sitemaker.umich.edu/group-based/home) for details) are free.

Several general use stat packages also have procedures for calculating power. SAS has **proc power**, which has a lot of features and is pretty nice. Stata has the **sampsi** command, as well as many user-written commands, including **fpower**, **powerreg** and **aipe** (written by our IDRE statistical consultants). Statistica has an add-on module for power analysis. There are also many programs online that are free.

For more advanced/complicated analyses, Mplus is a good choice. It will allow you to do Monte Carlo simulations, and there are some examples at http://www.statmodel.com/power.shtml (http://www.statmodel.com/power.shtml) and http://www.statmodel.com/ugexcerpts.shtml (http://www.statmodel.com/ugexcerpts.shtml) .

Most of the programs that we have mentioned do roughly the same things, so when selecting a power analysis program, the real issue is your comfort; all of the programs require you to provide the same kind of information.

Multiplicity

This issue of multiplicity arises when a researcher has more than one outcome of interest in a given study. While it is often good methodological practice to have more than one measure of the response variable of interest, additional response variables mean more statistical tests need to be conducted on the data set, and this leads to question of experimentwise alpha control. Returning to our example of drug A and placebo, if we have only one response variable, then only one *t* test is needed to test our hypothesis. However, if we have three measures of our response variable, we would want to do three *t* tests, hoping that each would show results in the same direction. The question is how to control the Type I error (AKA false alarm) rate. Most researchers are familiar with Bonferroni correction, which calls for dividing the prespecified alpha level (usually .05) by the number of tests to be conducted. In our example, we would have .05/3

= .0167. Hence, .0167 would be our new critical alpha level, and statistics with a p-value greater than .0167 would be classified as not statistically significant. It is well-known that the Bonferroni correction is very conservative; there are other ways of adjusting the alpha level.

Afterthoughts: A post-hoc power analysis

In general, just say "No!" to post-hoc analyses. There are many reasons, both mechanical and theoretical, why most researchers should not do post-hoc power analyses. Excellent summaries can be found in Hoenig and Heisey (2001) *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis* and Levine and Ensom (2001) *Post Hoc Power Analysis: An Idea Whose Time Has Passed?*. As Hoenig and Heisey show, power is mathematically directly related to the p-value; hence, calculating power once you know the p-value associated with a statistic adds no new information. Furthermore, as Levine and Ensom clearly explain, the logic underlying post-hoc power analysis is fundamentally flawed.

However, there are some things that you should look at after your study is completed. Have a look at the means and standard deviations of your variables and see how close they are (or are not) from the values that you used in the power analysis. Many researchers do a series of related studies, and this information can aid in making decisions in future research. For example, if you find that your outcome variable had a standard deviation of 7, and in your power analysis you were guessing it would have a standard deviation of 2, you may want to consider using a different measure that has less variance in your next study.

The point here is that in addition to answering your research question(s), your current research project can also assist with your next power analysis.

Conclusions

Conducting research is kind of like buying a car. While buying a car isn't the biggest purchase that you will make in your life, few of us enter into the process lightly. Rather, we consider a variety of things, such as need and cost, before making a purchase. You would do your research before you went and bought a car, because once you drove the car off the dealer's lot, there is nothing you can do about it if you realize this isn't the car that you need. Choosing the type of analysis is like choosing which kind of car to buy. The number of subjects is like your budget, and the model is like your expenses. You would never go buy a car without first having some idea about what the payments will be. This is like doing a power analysis to determine approximately how many subjects will be needed. Imagine signing the papers for your new Maserati only to find that the payments will be twice your monthly take-home pay. This is like wanting to do a multilevel model with a binary outcome, 10 predictors and lots of cross-level interactions and realizing that you can't do this with only 50 subjects. You don't have enough

“currency” to run that kind of model. You need to find a model that is “more in your price range.” If you had $530 a month budgeted for your new car, you probably wouldn't want exactly 530 in monthly payments. Rather you would want some “wiggle-room” in case something cost a little more than anticipated or you were running a little short on money that month. Likewise, if your power analysis says you need about 300 subjects, you wouldn’t want to collect data on exactly 300 subjects. You would want to collect data on 300 subjects plus a few, just to give yourself some “wiggle-room” just in case.

Don’t be afraid of what you don’t know. Get in there and try it BEFORE you collect your data. Correcting things is easy at this stage; after you collect your data, all you can do is damage control. If you are in a hurry to get a project done, perhaps the worst thing that you can do is start collecting data now and worry about the rest later. The project will take much longer if you do this than if you do what we are suggesting and do the power analysis and other planning steps. If you have everything all planned out, things will go much smoother and you will have fewer and/or less intense panic attacks. Of course, some thing unexpected will always happen, but it is unlikely to be as big of a problem. UCLA researchers are always welcome and strongly encouraged to come into our walk-in consulting (/schedule/statistical-consulting-schedule/) and discuss their research before they begin the project.

Power analysis = planning. You will want to plan not only for the test of your main hypothesis, but also for follow-up tests and tests of secondary hypotheses. You will want to make sure that “confirmation” checks will run as planned (for example, checking to see that interrater reliability was acceptable). If you intend to use imputation methods to address missing data issues, you will need to become familiar with the issues surrounding the particular procedure as well as including any additional variables in your data collection procedures. Part of your planning should also include a list of the statistical tests that you intend to run and consideration of any procedure to address alpha inflation issues that might be necessary.

The number output by any power analysis program is often just a starting point of thought more than a final answer to the question of how many subjects will be needed. As we have seen, you also need to consider the purpose of the study (coefficient different from 0, precise point estimate, replication), the type of statistical test that will be used (t-test versus maximum likelihood technique), the total number of statistical tests that will be performed on the data set, genearlizability from the sample to the population, and probably several other things as well.

The take-home message from this seminar is “do your research before you do your research.”

References

Anderson, N. H. (2001). Empirical Direction in Design and Analysis. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bausell, R. B. and Li, Y. (2002). Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences. Cambridge University Press, New York, New York.

Bickman, L., Editor. (2000). Research Design: Donald Campbell's Legacy, Volume 2. Thousand Oaks, CA: Sage Publications.

Bickman, L., Editor. (2000). Validity and Social Experimentation. Thousand Oaks, CA: Sage Publications.

Campbell, D. T. and Russo, M. J. (2001). Social Measurement. Thousand Oaks, CA: Sage Publications.

Campbell, D. T. and Stanley, J. C. (1963). Experimental and Quasi-experimental Designs for Research. Reprinted from *Handbook of Research on Teaching*. Palo Alto, CA: Houghton Mifflin Co.

Chen, P. and Popovich, P. M. (2002). Correlation: Parametric and Nonparametric Measures. Thousand Oaks, CA: Sage Publications.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, Second Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cook, T. D. and Campbell, D. T. Quasi-experimentation: Design and Analysis Issues for Field Settings. (1979). Palo Alto, CA: Houghton Mifflin Co.

Graham, J. W., Cumsille, P. E., and Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka and W. F. Velicer (Eds.), *Handbook of psychology* (Vol. 2, pp. 87-114). New York: Wiley.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26(3)*, 499-510.

Hoenig, J. M. and Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician, 55(1)*, 19-24.

Kelley, K and Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8(3)*, 305-321.

Keppel, G. and Wickens, T. D. (2004). Design and Analysis: A Researcher's Handbook, Fourth Edition. Pearson Prentice Hall: Upper Saddle River, New Jersey.

Kline, R. B. Beyond Significance (2004). Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. American Psychological Association: Washington, D.C.

Levine, M., and Ensom M. H. H. (2001). Post Hoc Power Analysis: An Idea Whose Time Has Passed? *Pharmacotherapy, 21(4)*, 405-409.

Lipsey, M. W. and Wilson, D. B. (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-analysis. *American Psychologist, 48(12)*, 1181-1209.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5(4)*, 434-458.

Maxwell, S. E. and Delany, H. D. (2004). Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Murphy, K. R. and Myors, B. (2004). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Publication Manual of the American Psychological Association, Fifth Edition. (2001). Washington, D.C.: American Psychological Association.

Sedlmeier, P. and Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin, 105(2)*, 309-316.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). Experimental and Quasi-experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin Co.

Stratton, I. M. and Neil, A. (2004). How to ensure your paper is rejected by the statistical reviewer. *Diabetic Medicine*, 22, 371-373.

Tversky, A. and Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin, 76(23)*, 105-110.

Webb, E., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (2000). Unobtrusive Measures, Revised Edition. Thousand Oaks, CA: Sage Publications.

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)

