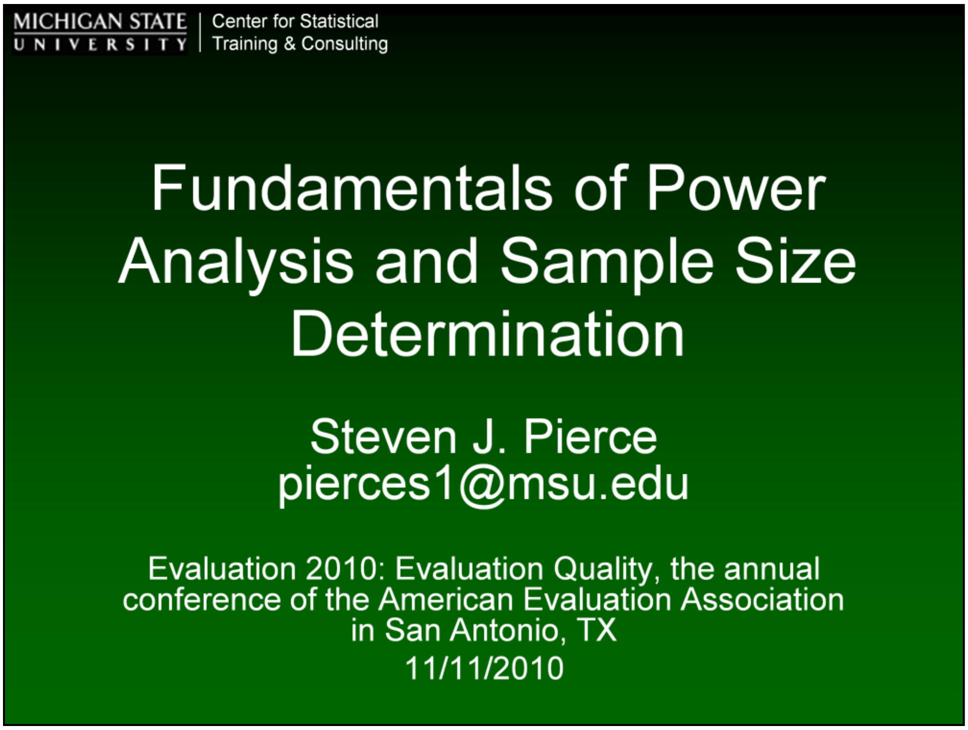
MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Citation

Pierce, S. J. (2010, November 10-13). *Fundamentals of power analysis and sample size determination.* Demonstration session presented at Evaluation 2010: Evaluation Quality, the annual conference of the American Evaluation Association, San Antonio, TX.

Abstract: In quantitative studies, statistical power (the probability of detecting an effect that actually exists) is closely tied to sample size. Evaluators can use power analysis to plan what sample size should be targeted during data collection to make best use of limited evaluation resources. This introductory session will cover the fundamental concepts involved in using power analysis and describe how power analysis can be used to improve the quality of a quantitative evaluation study. It will define key terms, explain why power analysis is important, and then discuss practical issues such as how to pick a power analysis method that matches your hypotheses, how to come up with reasonable numbers to plug into power analysis formulas, and why it is important to examine how sensitive the results are to your assumptions. Some examples will be presented, and software tools and other resources will be recommended.

Relevance: Power analysis is currently the best available method for planning the sample size required for conducting a high-quality, quantitative study. Because evaluators have an ethical responsibility to ensure that evaluation resources are used wisely, they need to be aware of the potential problems associated with both inadequate and excessive sample sizes. Studies that don't collect enough data are at serious risk of failing to adequately test the hypotheses that motivate the evaluation, while studies that collect more data than necessary may be diverting resources that could be allocated to other, more productive uses. Educating evaluators about power analysis has the potential to improve evaluation quality because it may help them to design efficient studies that collect enough data to meet the technical standards applicable to the guiding principle of systematic inquiry, but that do not waste their clients' resources on excessive data collection. Audience members will learn what kinds of information are required to do power analysis and what kinds of assumptions must be made. They will receive concrete suggestions on how to use previous literature, pilot data, and subject matter knowledge to inform the decisions made during a power analysis. In addition to seeing some examples based on simple evaluation designs, the audience will be given an annotated resource list with suggested textbooks, articles, and software that will allow them to learn more on their own.



Good afternoon. I'm Steve Pierce and I'll be introducing you to the fundamental concepts and issues in using power analysis to determine the sample size you might need for a quantitative study. I'll be posting electronic copies of the slides (complete with my speaker notes) on my website and on AEA's public e-Library within the next day or two, along with a list of recommended resources. You can also get these materials from me via the e-mail address listed in the conference program.

I'll be focusing on key principles and concepts because the way you do a power analysis depends on the statistical methods you will use to analyze the data. There are far more statistical methods than we have time to discuss, so the examples will illustrate how to do power analyses for 2 simple, widely used statistical tests. Fortunately, the plethora of specific power analysis formulas all rest on some common foundations. If you understand those, it will be much easier to figure out how to use the formulas associated with more sophisticated statistical methods.

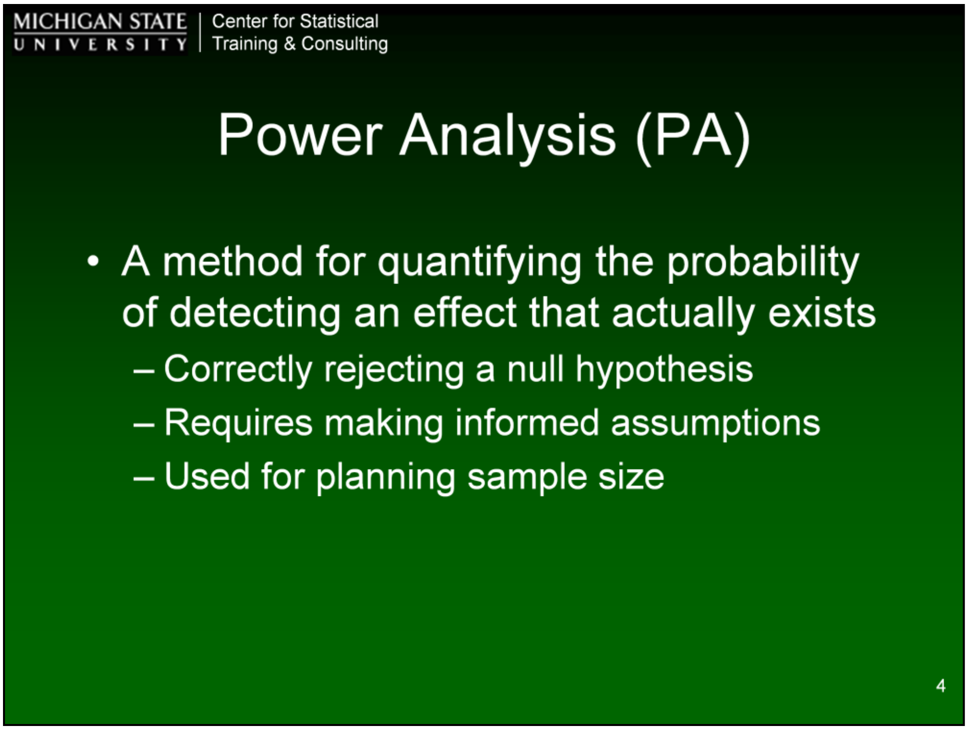
I want this session to be useful regardless of whether you will personally be running power analyses, or you will work with a statistician who will do them for you. In the latter case, the statistician will need information from you to run a good power analysis. Hopefully, this session will help you think clearly about the key issues and be prepared to answer the statistician's questions.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Outline

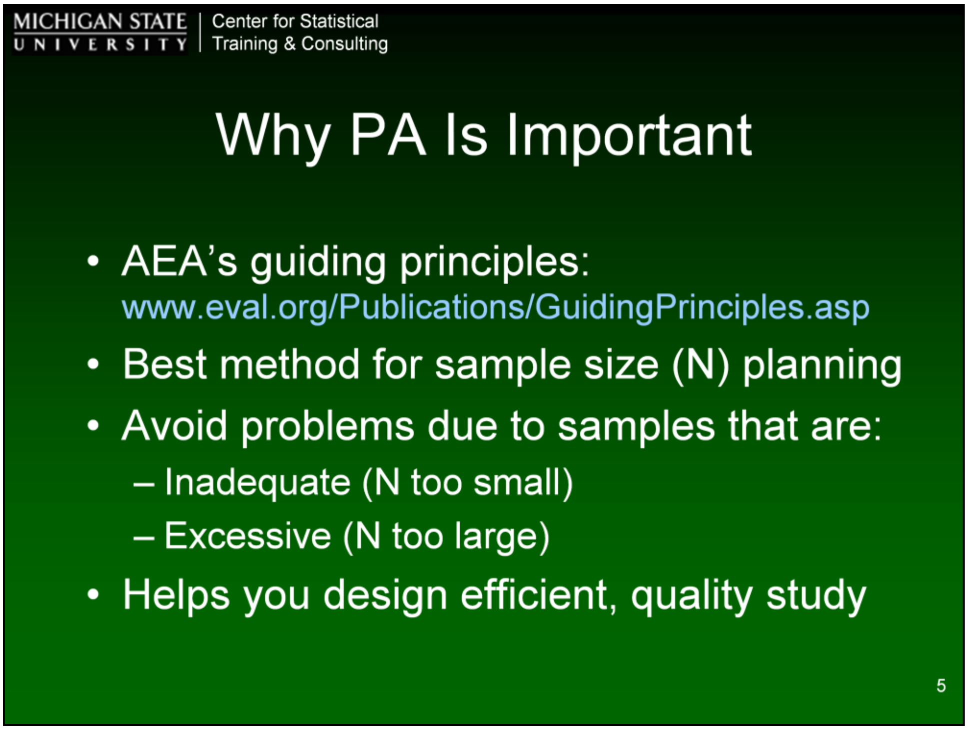
- Define power analysis
- Why PA is important
- Statistical inference
- What affects power
- Ways to apply PA
- Conventions for α & β
- Why analysis methods matter
- Dissecting effect size measures
- Present 2 examples
- Practical advice
- Questions & discussion

3



So, let's get started by defining power analysis. In simple terms, it's just a method for quantifying the probability that you will be able to detect an effect that really exists. So, say you're evaluating an educational intervention where you expect the experimental group to perform better than the control group on math tests. Power analysis can tell you how likely you are to detect a difference in math scores between the two groups assuming that the program really works. So, because the focus is on situations where the effect of interest really exists, we are saying that we want to know how likely we are to correctly reject a null hypothesis that says there is no effect.

To do that, we need to make some informed assumptions about how big an effect is worth detecting and how large your sample is going to be. I'll be talking more about what assumptions you might need to make and how you use prior research, theory, and/or pilot data to inform your assumptions. For now, just note that you'll have to make assumptions and that the quality of the power analysis depends very much on how well informed they are. Ultimately, you should use power analysis results to plan what sample size you will need for your evaluation study. But, you also need to look at how power changes as you change the sample size you might use or the other assumptions in the power analysis. So, that's a very fast intro to what power analysis is. Now let's talk about why it's important to do one before you start a new study.



I assume you are already familiar with AEA's guiding principles for evaluators (www.eval.org/Publications/GuidingPrinciples.asp). If not, I urge you to review those soon because they are directly relevant to why power analysis is important. For example, power analysis is the best available method for planning the sample size required to conduct a high-quality, quantitative study, so doing one helps you meet the expectations in the principle of Systematic Inquiry.

Under the principle of Integrity & Honesty, we are obligated to discuss the actual and potential limitations of our methods with clients. Studies that don't collect enough data are at serious risk of failing to adequately test the hypotheses that motivate the evaluation. You're in a better position to discuss whether the planned sample size will be adequate if you've done a good power analysis.

On the other hand, studies that collect too much data may be unnecessarily burdensome to evaluation participants and thereby fail to meet the principle of Respect for Persons. They may also be diverting resources that could be allocated to other, more productive uses. Surely the principle of Responsibility for General & Public Welfare encompasses an ethical responsibility to ensure that evaluation resources are used wisely and not wasted on excessive data collection. Power analysis can help you to design efficient, high-quality studies that collect enough data to meet the technical standards applicable to the guiding principle of systematic inquiry, without over-doing it.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Statistical Inference in NHST

Decision → **Reality ↓**	**Fail to Reject H_0**	**Reject H_0**
H_0 is True	Correct There's no effect $P = 1 - \alpha$	Type I Error Falsely conclude there's an effect $P = \alpha$
H_0 is False	Type II Error Missed real effect $P = \beta$	Correct Found real effect $P = 1 - \beta$ = Power

H_0 = null hypothesis, P = probability

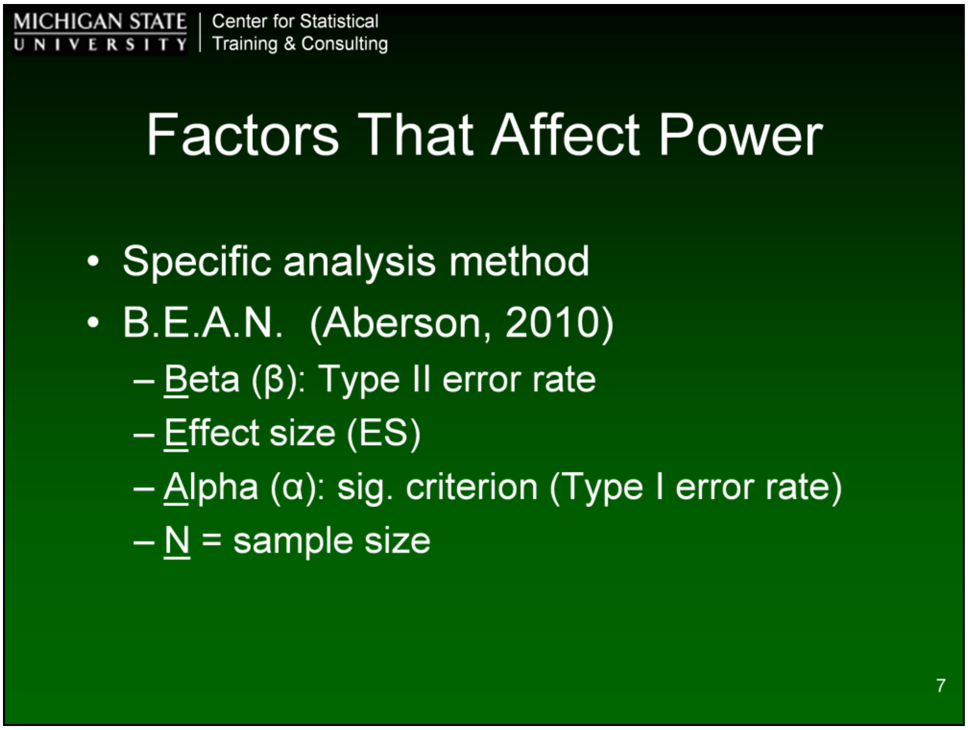
6

Power analysis is closely tied to a framework for making statistical inferences that is often called null-hypothesis significance testing, which is abbreviated as NHST. Here's how that framework works. You start with a null hypothesis (which is usually labeled H0) that says there is no effect. For example, it might state that there's no difference between the mean test scores of the intervention and control groups. NHST starts from the assumption that H0 is true and then requires you to see compelling evidence that the data you've observed are very unlikely to have occurred in that situation before you decide to reject the null hypothesis.

Now, in reality, H0 is either true (there really is no difference), or false (the means are actually different). Unfortunately, we have to analyze data, then decide whether we are convinced there really is an effect (causing us to reject H0) or whether the evidence is insufficient to convince us of that (we fail to reject H0). That is a statistical inference based on the available data. Of course, depending on how our decision matches up with the real state of things, we can make either of two different kinds of errors.

If H0 is actually true, we can only make what is called a Type I error by falsely concluding there's an effect when there really isn't. We don't want to do that very often, so we usually choose methods that will give us a small probability of doing that. The typical α = 0.05 criterion you've all heard about in your statistics classes is just saying you only want to make that kind of error 5% of the time.

If H0 is actually false, we can only make a Type II error, which means that we will fail to detect a real effect. The probability of making such an error is called β, so power is just the probability that you will actually detect a real effect. If you have a 20% chance of failing to detect the effect, then you must have an 80% chance of correctly detecting it.



So, what actually affect the statistical power for a study? First and foremost, the specific analysis method you'll be using (which should be influenced by your research design) is a key factor because it controls the actual formulas you need to use to do the power analysis. I'll cover that in more depth later, by using some concrete examples.

Second, the acronym BEAN nicely captures the four factors that affect power once you've selected an analysis method.

The B in BEAN stands for β, which is the Type II error rate. Subtracting β from 1 gives you the expected power, so if you know β, then you know how much power your study will have. The lower your β error rate gets, the higher the power your study will have.

The E stands for effect size. The larger the effect size you're looking for, the easier it is to detect it. That means if all else is equal, you will always have more power to detect large effects than you do to detect small effects. Since measuring effect size is one of the most poorly understood parts of power analysis, I'll focus a lot on that in the examples.

The A stands for the significance criterion α, which is the Type I error rate you are willing to risk. If you're willing to be more liberal about how often you could falsely conclude there's an effect, you will have more power. That lowers the bar for how much evidence you need before you decide there is a significant effect. The convention in the social sciences is to set α = .05. There are situations when you might want to be more conservative by setting it lower (say α = .01), making it less likely you'll claim there's an effect when there really isn't, but the cost of doing that is decreased power. Conversely, you may want to tradeoff a higher risk of a Type I error in order to get more power to detect effects.

Finally, the N in BEAN stands for the sample size you uses in the study. Larger samples always give you more power to detect real effects than do small samples, if everything else is the same between two studies.

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences. New York, NY: Routledge.*

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Applying Power Analysis

Given:	You Can Estimate:
Typical (Common)	
1. β, ES, α	Required N
2. ES, α, N	Expected β (power)
Atypical (Rare)	
3. α, N, β	Minimum detectable ES
4. β, ES, N	Required α

8

The factors identified in the BEAN acronym are inter-related. If we know values for any 3 of them, we can use those relationships to calculate the other one. The first two rows in this table show you the most common applications of power analysis which are the focus of this talk because they are really the most useful in planning evaluations, while the last two rows show relatively rare applications.

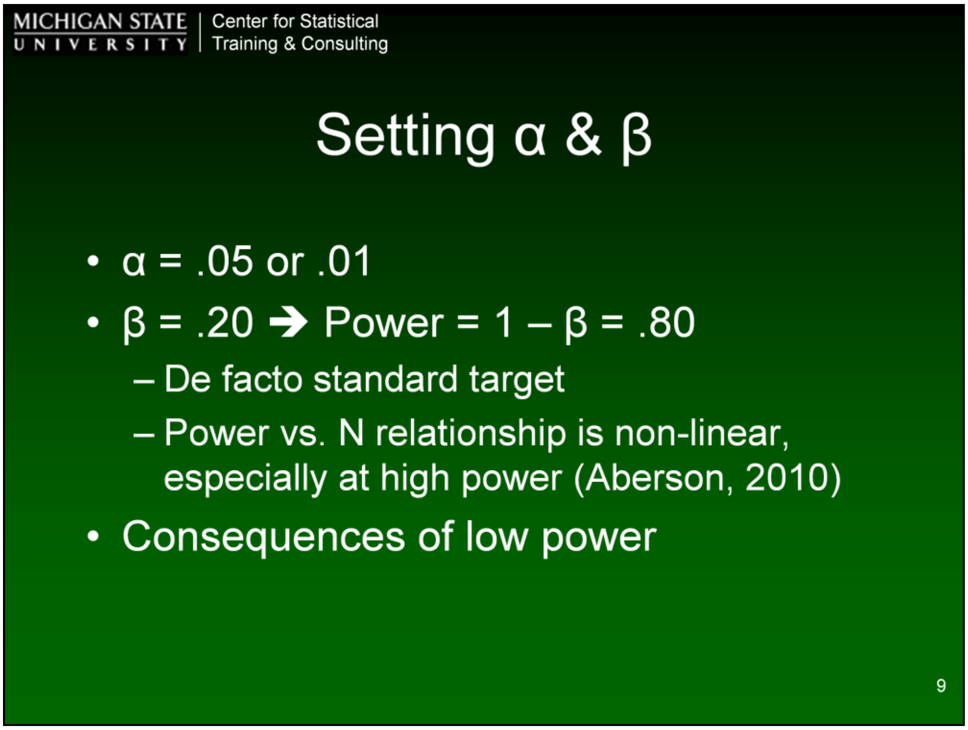
In row 1, you know how much power you want, the ES you want to be able to detect, and the α error rate you're willing to accept, then you use power analysis formulas to calculate how large a sample you need. This is a good way to plan the sample size you will need for a study.

In row 2, you may have a constraint on the sample size (because of the data collection cost per person), but you know the ES you want to detect and the α error rate you're willing to accept, so you use power analysis to figure out how much power you would really have before you commit to doing the study. If it is too low, you may decide that it's not worth doing the study.

In row 3, you may have a situation where you know the α error rate, power, and the sample size, but want to know how large an effect you can reliably detect. This is a less common application because it doesn't help you plan a new study. It's mostly used for post-hoc power analyses, which statisticians have criticized as a seriously flawed endeavor. We don't have time to discuss that controversy, but you can read about it in the literature on power analysis.

In row 4, you may already know the desired power, the effect size, and the sample size, but you need to know how large your type I error rate must be to achieve that level of power given the other constraints. This is also a pretty rare application.

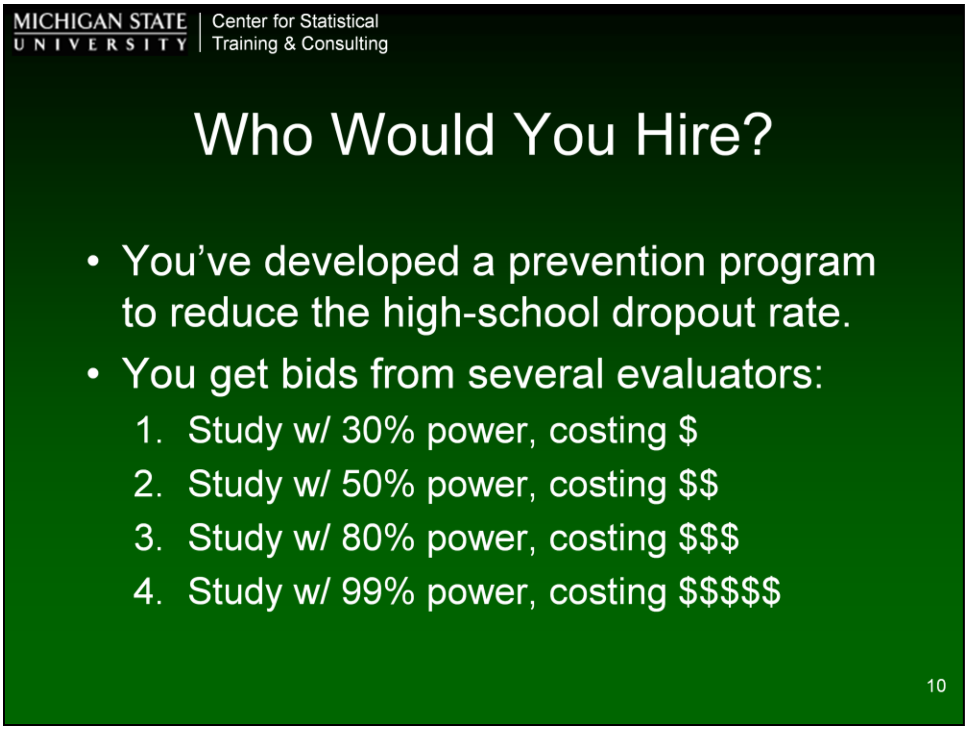
So, you can see that we have several input factors that we need to set when doing a power analysis, one of which is usually the effect size. I'm going to talk about that part a little later because I want to cover how we handle α and β first.



Fortunately α and β are fairly easy to set because we have conventions in the social sciences that guide us in choosing desirable levels for the α and β error rates. We usually set α at either .05 or .01 so that we have a fairly low probability of falsely deciding there's an effect when there isn't. That's such a common practice that most reviewers or statisticians would only question you if you deviated from that practice. If you do that, you want to be ready to defend that choice (especially if you are setting it higher).

We also have some de facto standards for setting the β error rate (and hence power) as well. You want high power because that means you have a reasonable chance of detecting a real effect. To get high power, the β error rate has to be low. A 20% error rate will give you 80% power. That's the typical target recommended in the social science literature. Now, there are times when you might want more power, but it's important to realize that the relationship between power and sample size is not really linear. Once you get above about 80% power, you usually need larger increases in sample size to achieve small increases in power, especially for small effect sizes. However, if the marginal cost of additional data collection is low (as it might be with a web-based survey), it might well be worth aiming for 90-95% power. It might also be important to have really high power if there are major consequences to committing a Type II error by failing to detect a real effect. For example, say you are testing whether or not a program is associated with an side effect that has very serious adverse consequences for people. You probably want really high power to detect such an outcome.

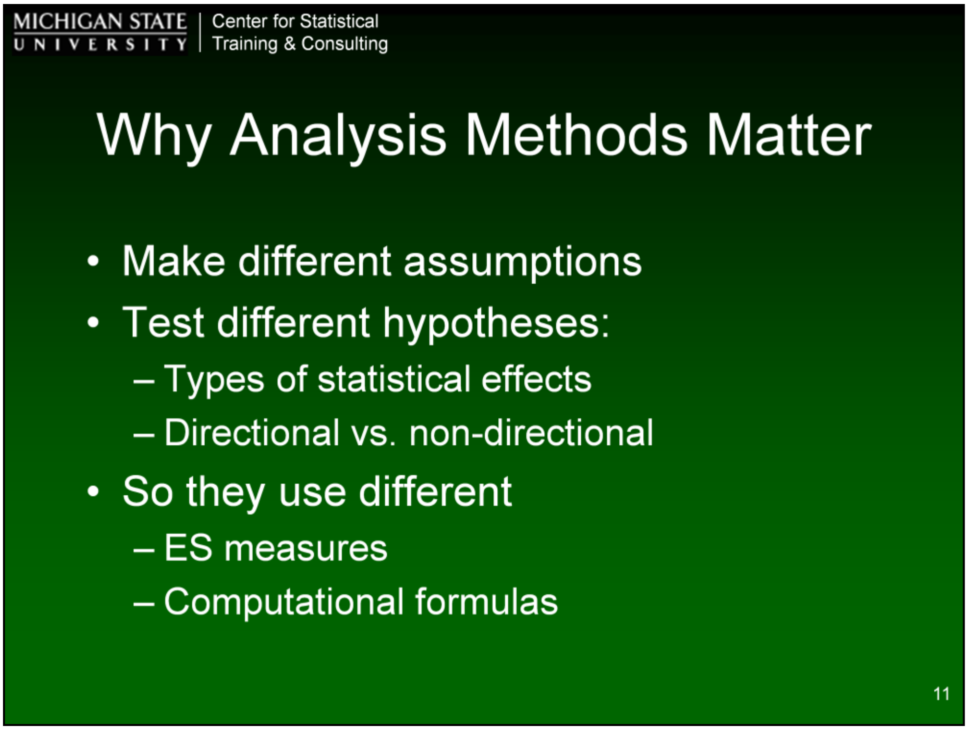
If your power is too low, you're not really subjecting your hypothesis to a fair test: You're stacking the odds against actually detecting an effect.



So, let's spend a moment thinking about the desired level of power in a real-world sort of situation. Suppose you've developed a prevention program to reduce high-school dropout rates, and now you want to have an evaluation done so you see if it really works as well as you hope. You'd like to be able to show the world that it works and then disseminate it widely. You put out a call for proposals and get bids from several evaluators. Being an educated program developer, you study the proposals and conclude that the study designs differ dramatically in terms of both statistical power and the costs associated with data collection, but are all otherwise acceptable.

Which evaluator would you hire?

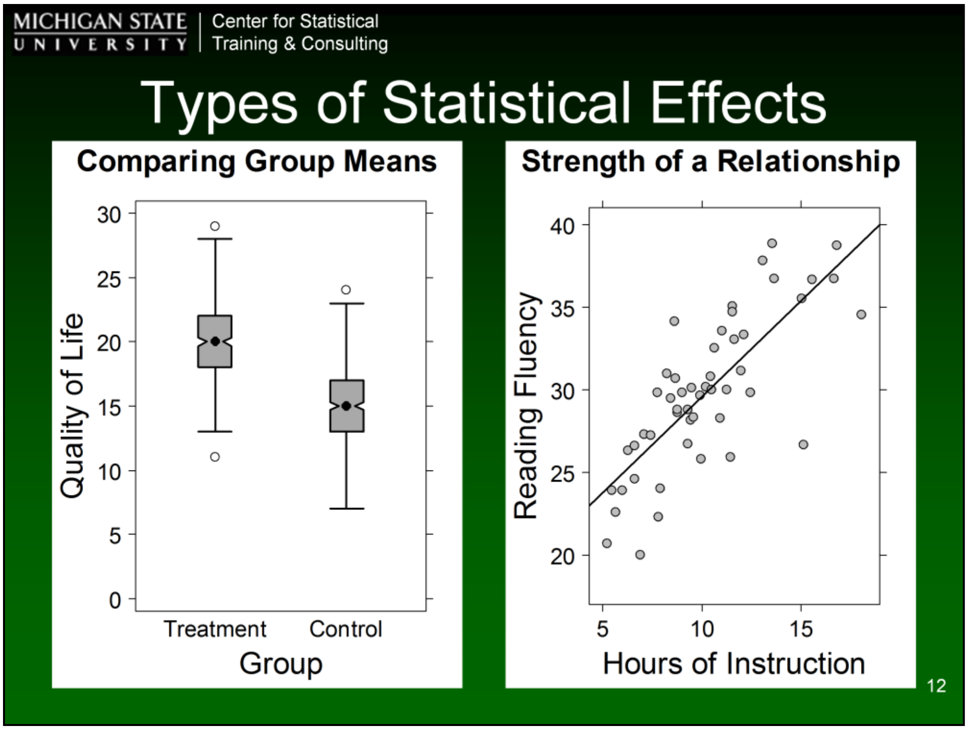
[After discussion] OK. So let's get back to talking about the other factors you need to understand in order to use power analyses to plan a study.



Now I'd like to elaborate on why the analysis methods you plan to use are important when you want to do a power analysis. Researchers have developed a tremendous array of statistical methods for analyzing data over the years. That has important implications for power analysis because the various statistical models make different assumptions about things like whether the data in question are categorical or continuous, about the distributions of continuous variables, and so on.

Analysis methods usually make different assumptions because they aim to test different hypotheses. For example, the hypotheses depend on the type of statistical effect being examined, and on whether or not the researcher wants to perform a directional hypothesis test. A directional test might ask whether a new program is superior to an existing program with respect to producing some outcome. That's a directional test because it only asks whether the new program is better – and asking the question that way indicates that it doesn't matter if the new program is actually worse than the old one. That's sensible if the new program is sufficiently more expensive that it will only be adopted if it is demonstrably better then the current one. If it's equal to the old one, or actually worse, then the extra cost alone would prevent the new program from being implemented.

Because of the different assumptions and hypotheses involved, different statistical models rely on different measures of effect size and computational formulas, so the corresponding power analyses must do so as well. To really understand effect sizes, we have to first understand a bit more about different types of statistical effects. So, let's take a closer look at that concept.

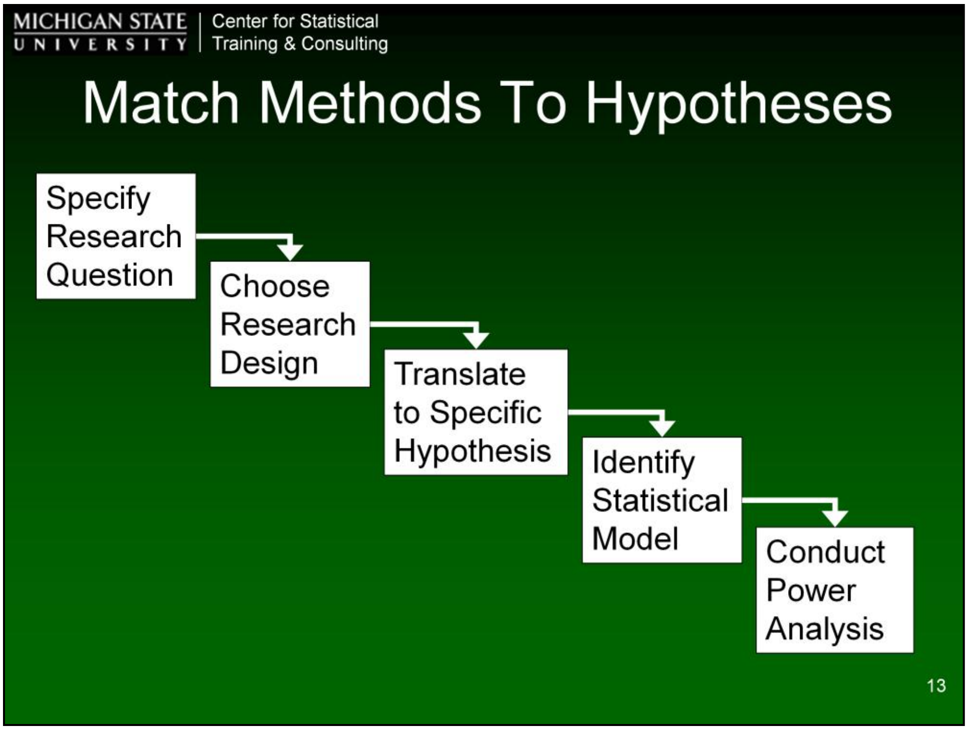


Here are graphs illustrating two different types of statistical effects we might be interested in for evaluation studies. The analysis method you will use depends on what kind of effect you want to test.

In the first example, you might want to compare groups of participants by looking at whether they tend to be similar or different on some outcome measure. So, in the graph on the left you can see boxplots showing the distributions of quality of life scores for two different groups: a treatment group and a control group. Here, you'd usually be testing whether the means for the two groups are equal. So, the statistical effect of interest is a difference between two group means.

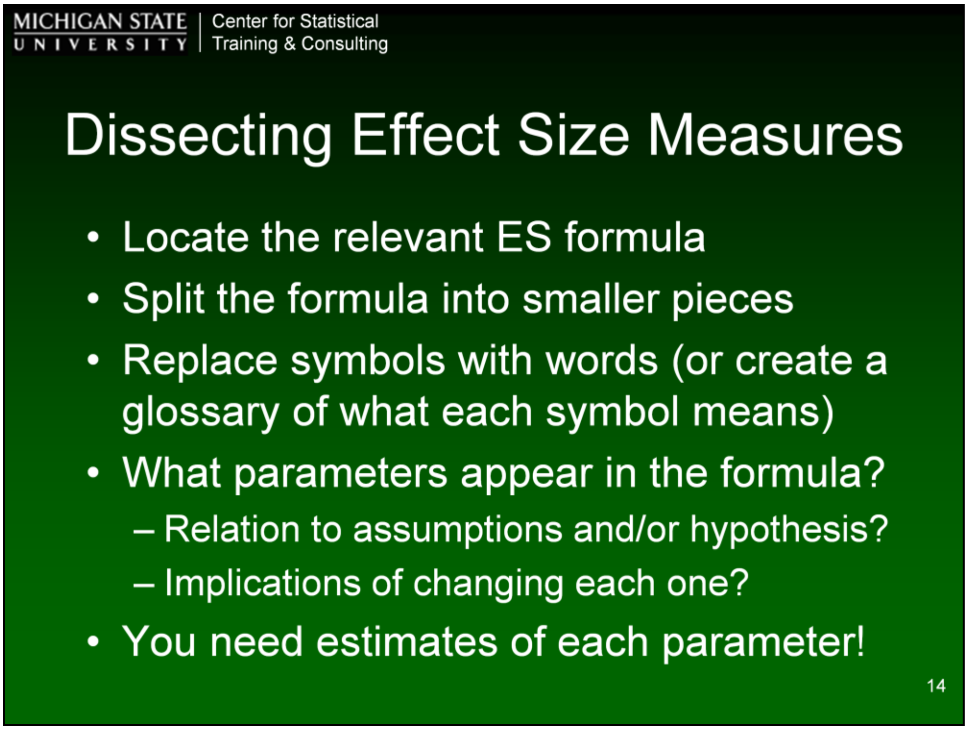
In the second example, you might want to test whether there is a statistical relationship between two different variables. In the graph on the right, we're interested in whether there is a relationship between hours of instruction and reading fluency. So here the effect of interest has to quantify the strength of that relationship.

If we want to use the simplest and most straightforward methods, testing these two kinds of effects requires using different statistical tests because they examine different hypotheses. So, it stands to reason that they will use different measures of effect size, and thus different power analysis formulas. My examples will cover power analyses for these two types of effects.



So, that's why I want to emphasize how important it is that you properly match your methods to your hypotheses. You really can't do a very good power analysis until you've specified your research question and research design. Once you do that, you need to translate your question into a specific hypothesis that can be tested to answer your question. Then, you need to identify a statistical method for analyzing the data that is actually designed to test the kind of hypothesis you've generated.

It's not appropriate to use a power analysis designed for testing correlations to figure out how large a sample you need to detect a difference between group means. Those are different types of effects, so they rely on different statistical methods. You need to know what kind of analysis you'll be doing so you can tell which measure of effect size and which power analysis formulas are appropriate for your study.



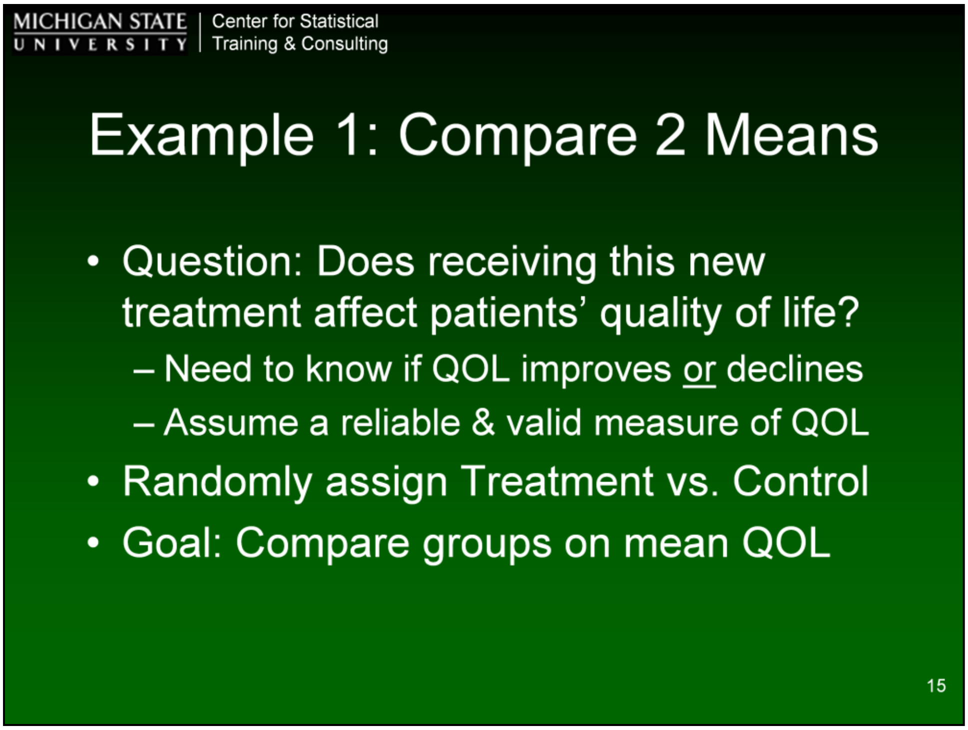
Now we can finally start talking about effect sizes, which are really at the heart of power analysis. ES measures are usually a little abstract, so they're hard to understand unless you translate them into more familiar terms. Here are some strategies I use for that.

Obviously, you have to locate the formula for the effect size measure used in the statistical model you plan to apply. You can usually look it up in a textbook, a journal article, or the help system of your power analysis software.

Once you have the formula, try dissecting it. Split it up into smaller pieces that you can understand more easily. Often, ES measures have multiple input parameters. They're also often expressed as fractions or have parts that are fractions. Pay attention to whether a parameter appears in a numerator vs. a denominator. That tells you something useful about how changing its value affects the resulting effect size.

Try to identify what the all parameters are and what each really represents. Re-write the equation in words instead of symbols, or create a glossary telling you what each symbol means. For example, the formula might include a pair of group means, or an expected correlation. Identifying these pieces allows you to better understand how the ES formula is related to the assumptions of the statistical model and how it relates to the hypothesis being tested. It also makes it easier for you to see what kinds of values are meaningful for each parameter and what happens if you change their values.

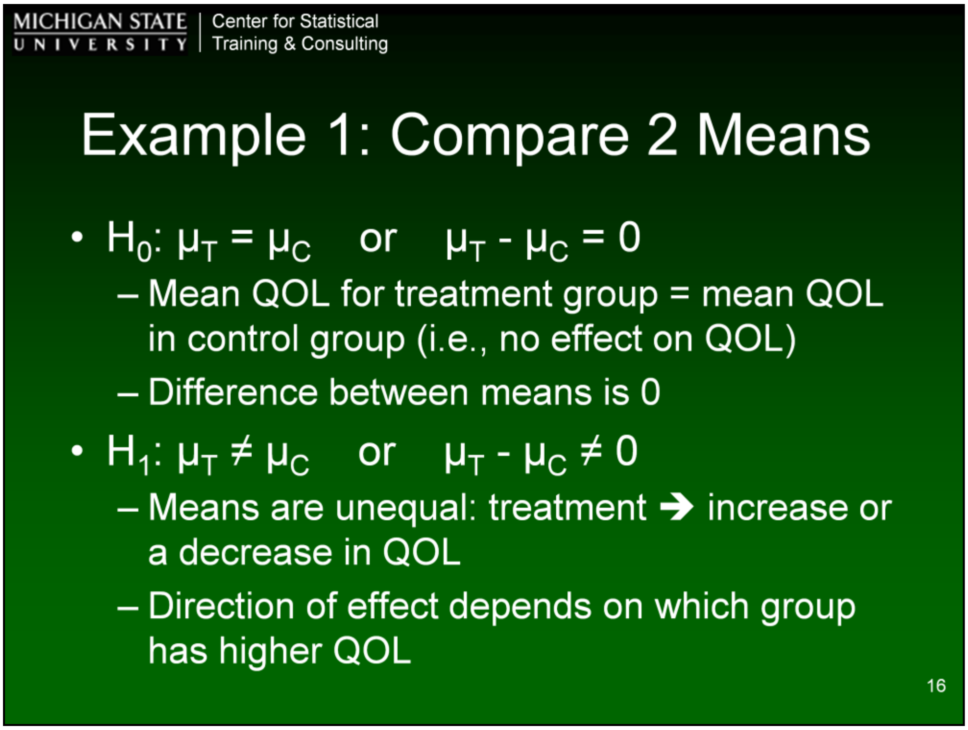
Ultimately, you need an estimate of each input parameter in the ES formula. Now, my experience has been that to really get your head wrapped around effect sizes, you need concrete examples, so that's what we're going to cover next. As I go through them, look for how these strategies apply in the examples.



My examples rely on very simple research designs and statistical analyses. More complex research designs and analysis approaches would certainly be worthwhile in real evaluations, but I chose these simple approaches so we can focus on the most fundamental concepts in power analysis. It really helps to understand the simplest cases well before you try to apply power analysis to more complex studies.

In this example, let's assume you're trying to evaluate whether or not a new treatment affects patients' quality of life. You need to know whether it improves quality of life (as hoped), but since it's a new treatment there's also a chance it may actually decrease quality of life compared to not getting the treatment. You need to know if it makes any difference in patients' quality of life, either positive or negative. Let's assume that you've already got a good measure of quality of life.

The simplest approach would be to randomly assign patients to either a treatment group or a control group, then measure QOL for all the patients after treating the experimental group. You can then compare the mean scores for the two groups to answer the question motivating the study.



With that situation, you can express the null hypothesis (H0) in very simple terms: You expect the mean QOL for the treatment and control groups to be equal. To put it another way, you expect the difference between the two means to be zero.

The alternative to that hypothesis (H1) is that the means are not equal, or that the difference between them is not equal to zero. In that case, then the sign of the difference (either positive or negative) indicates the direction of the difference, which depends on whether the treatment group has a higher or lower mean than the control group.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Example 1: Compare 2 Means

- Analysis: 2-tailed, independent t-test
  - Normal distribution for QOL scores
  - Equal variances for QOL across groups
  - Equal sample sizes across groups ($N = 2n_j$)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}}, df = N - 2$$

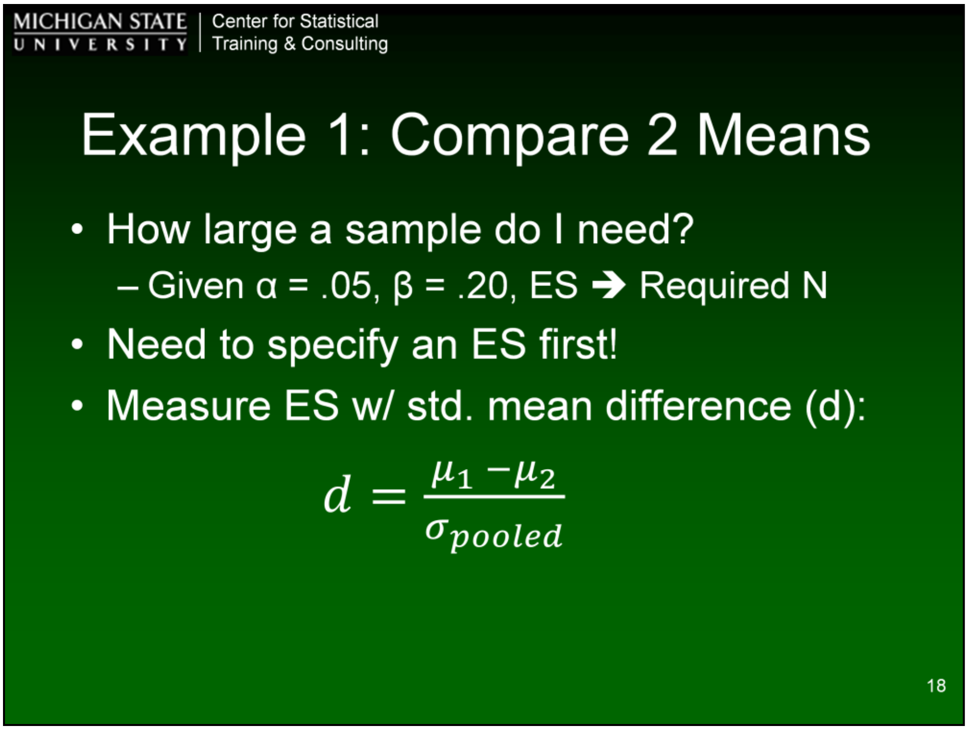
17

This as a hypothesis that can be tested with a two-tailed independent groups t-test. So, knowing that, we can plan to use that test to analyze the data from the study. That also tells us which set of formulas we need to use to do a power analysis.

To explain how the power analysis works, I just want to remind you of a couple key features of the t-test. First, it assumes that the outcome (QOL scores) follows a normal distribution. Second, to keep things as simple as possible, let's focus on the simplest type of t-test, which assumes the two groups will have both equal variances on QOL and equal sample sizes.

If the groups are equal in size, the total sample size N = 2 times nj (the sample size per group).

Here's the formula for the t-test. You can see that it basically looks at a ratio where the difference between the means is divided by the standard deviation of the difference between the means.



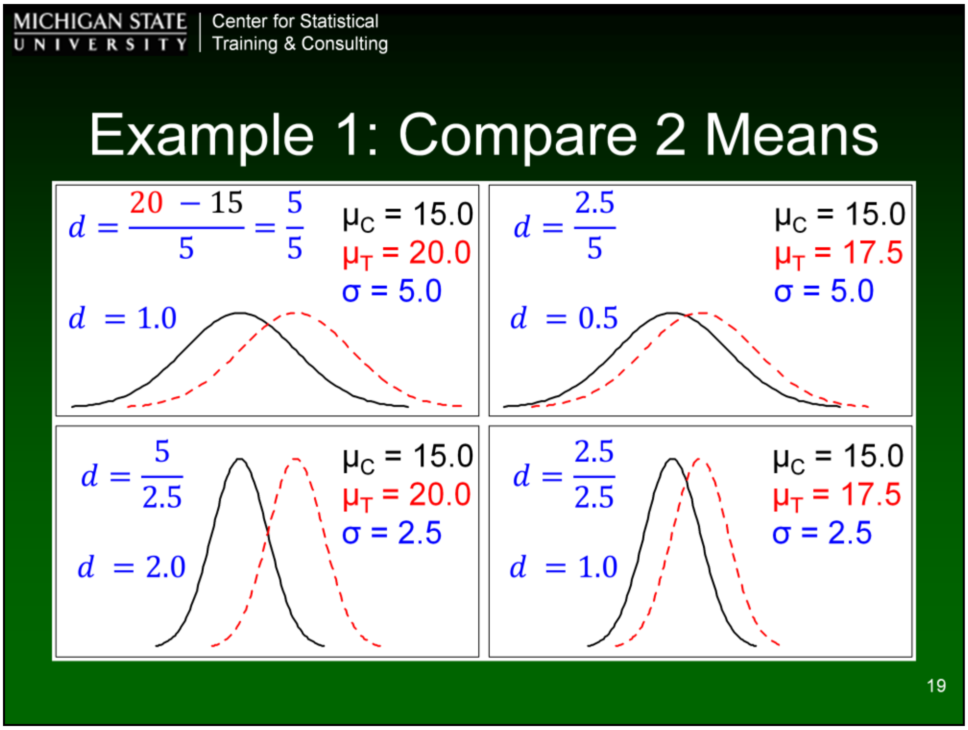
Let's say you want to use the most common type of power analysis, where the question is just "how large a sample do I need to test whether the treatment affects patients' QOL? ".

To calculate the required sample size, we need to pick values of alpha, beta, and effect size. Let's stick with the conventional values of alpha = .05 and beta = .20, which means we want to have 80% power to detect an effect of the treatment if it actually exists.

But wait – we still need to specify the effect size! Since how we measure that depends on the type of analysis we'll be doing, we look up the t-test in a good stats book and find that the relevant measure is called d and that it is calculated by dividing the difference between the means by the pooled standard deviation. It's essentially a standardized mean difference expressed in units of standard deviations. The value of d will get larger with bigger differences between the two means, but it will shrink if the standard deviation increases.

The numerator here ties directly to the hypothesis we want to test. Both the numerator and the denominator are tied to the assumption that the scores all come from normal distributions with the same variance. Because you only need a mean and a variance to describe a normal distribution, if you're assuming the variances are the same, the only way the two groups can differ is to have different means. But we want to scale that difference in relation to the amount of variability in the data.

If you have some idea of what the means for the two groups might be and what the standard deviation is, then you can specify an effect size and use that to finish calculating the required sample size.



Let's say we know that the mean QOL score = 15 with a SD = 5 points in previous research. That tells you what to expect for the control group (shown with solid black curves here). Since we're assuming normal distributions and equal variances, the only other number you need to calculate the effect size is the expected mean for the treatment group. On the upper left panel here, I've assumed the treatment group (shown as a dashed red curve) has a mean of 20 (5 points higher than the control group). That means the effect size is 5/5 = 1. On the upper right, I assumed the treatment group has a mean = 17.5, just 2.5 points higher than the control group, yielding an effect size of 0.5. Smaller differences between the means make the two distributions overlap more, making it harder to argue they are different. So, the effect size gets smaller as the difference shrinks.

On the bottom, I've changed to assuming that the SD = 2.5 points. On the left, the control group still has mean = 15, and the treatment group still has mean = 20, for a difference of 5 points. But, now the smaller SD makes the effect size much larger: 5/2.5 = 2. See how the distributions are narrower and overlap much less when we decrease the SD? In the lower right panel, reducing the difference in the means to 2.5 points once again produces an effect size of 1 (just like the upper left panel) because the difference in the means is 1 SD in size.

To continue walking you through the example, let's see how many people would be required to detect an effect size of 0.5 as shown in the upper right panel.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Example 1: Compare 2 Means

- α & β affect the noncentrality parameter (δ):

$$\delta = d\sqrt{\frac{n_j}{2}} = t_{critical} - t_{power}$$

$$\delta \cong z_{critical} - z_{power} = 1.96 - (-0.84)$$

$\delta \cong 2.80$ if α = .05 & β = .20

20

Now that you have an effect size, you need to know how alpha and beta get represented in the power analysis calculations. That gets captured by a concept called the noncentrality parameter, which is labeled with the symbol delta. When the null hypothesis is really true, the t-distribution is a symmetrical bell-shaped curve centered around the value t = 0. However, when it is false, the t-statistic follows an alternative, non-central distribution that is asymmetrical and somewhat skewed bell curve with the peak shifted toward one side. Delta just represents the distance between the centers of these two distributions. When it is very small and close to zero, the two distributions overlap a lot. As delta increases, they overlap less and less.

Delta is related to the effect size and to sample size as you can see here, but for this application of power analysis where we are trying to calculate required sample size, we take advantage of the fact that delta is also equal to the sum of two t-statistics: the critical t-value above which you would reject the null hypothesis (based on alpha), and the t value above which you find the percentage of the non-central t-distribution corresponding to your desired level of power (based on beta).

Because the t-distribution depends on sample size (which you don’t know yet), but is very close to the z-distribution when sample size gets above about 10, most of the time we just approximate delta by swapping in z-scores for the t-scores. So, here if you want alpha = .05 and beta = .20, delta = 2.80.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Example 1: Compare 2 Means

If α = .05, β = .20, and d = 2.5/5.0 = 0.5:

$$n_j = \frac{2\delta^2}{d^2} = \frac{2(2.80^2)}{0.5^2} = \frac{2(7.84)}{.25}$$

$$n_j = \frac{15.68}{.25} = 62.72$$

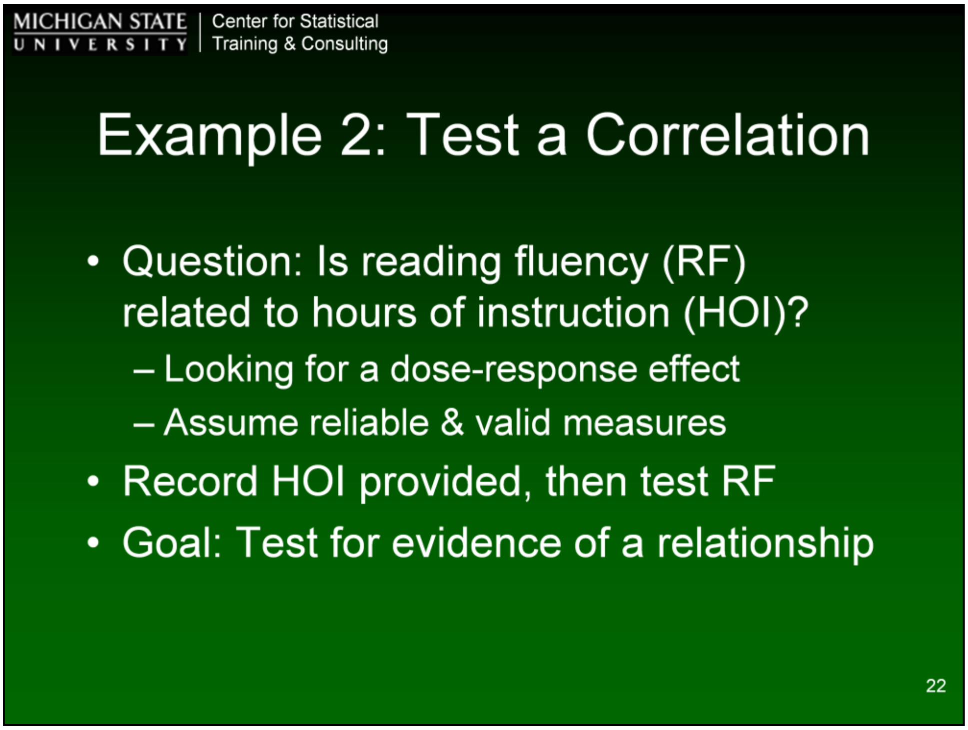
Need nj ≥ 63 per group for 80% power

21

Now we finally have all the numbers we need to calculate the required sample size. Here's the last step.

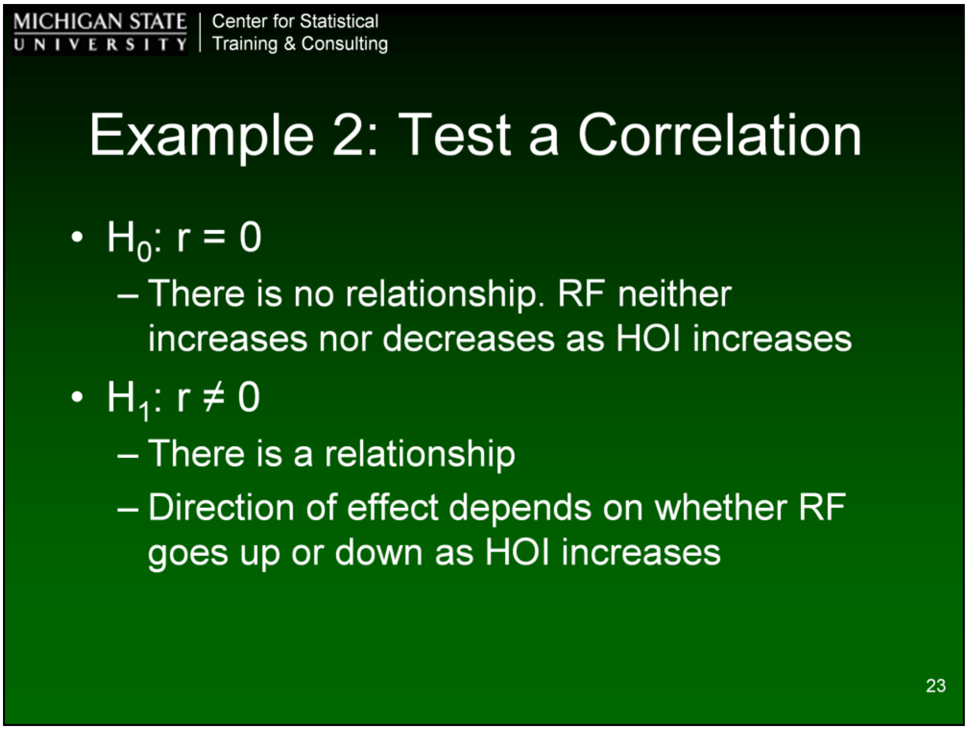
You can see here that the sample size per group (nj) equals 2 times the square of delta, all divided by the square of the effect size. First we swap in the numbers that represent our assumptions, then after we simplify the resulting equation.

We see that you need at least 63 people per group (a total of 126 people) in order to have an 80% chance of detecting a difference of half a standard deviation between the means of the treatment and control groups.



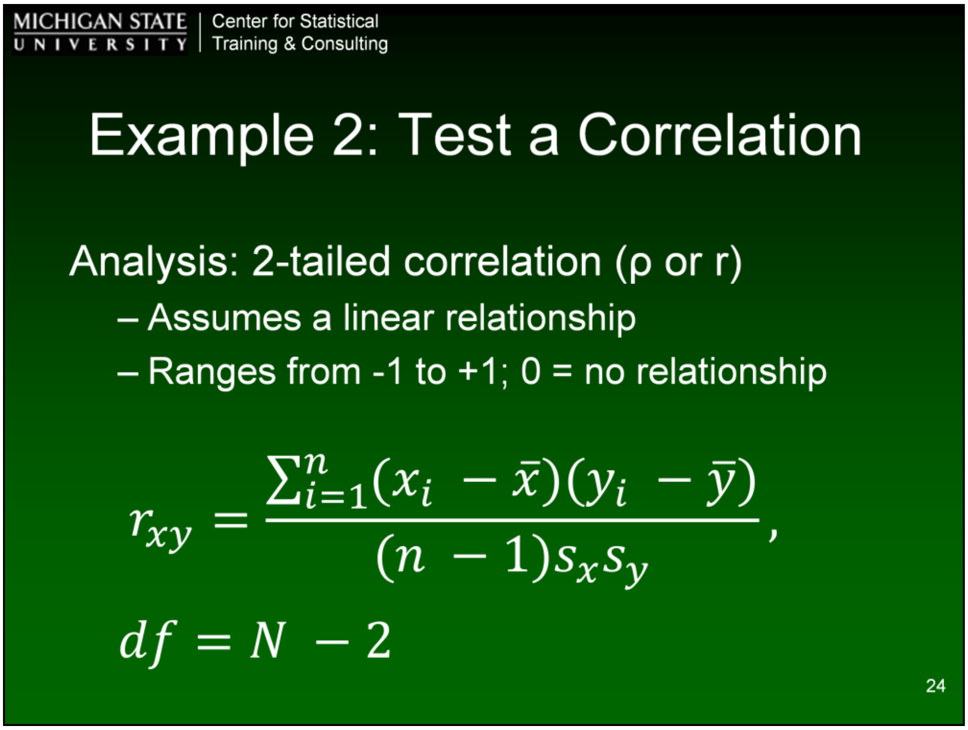
My second example involves testing for a different kind of statistical effect: one that measures the relationship between two variables. Let's say you're evaluating an educational program intended to help students become better at reading. But, since it is a voluntary after-school program, the participants don't all spend the same amount of time getting the intervention, so you can't say they've been equally exposed to the program.

One way to do that evaluation would be to look for evidence of a dose-response relationship between hours of instruction received by participants and an outcome like reading fluency. Assuming you have reliable and valid ways to measure both variables, the simplest approach to this is to record how much time individual participants spent getting instruction through the program, then give them each a test of reading fluency. After that, you can examine the data for evidence of a relationship between the two variables.



With that situation, you can express the null hypothesis (H0) in very simple terms: You expect the correlation to be zero, such that RF does not change systematically as HOI increases.

The alternative to that hypothesis is that the correlation is not zero. The sign of the correlation indicates the direction of the relationship: A positive correlation would indicate that as HOI increases, so does RF. That's what we would hope to see as evidence the program is beneficial. In contrast, a negative correlation would tell us that as HOI increases, RF actually decreases. That would tell us that the intervention is actually impairing the reading skills of the participants.



This as a hypothesis that can be tested with a Pearson correlation coefficient. We use the Greek symbol rho (ρ) for the population correlation, and the letter r for the sample correlation.

Here’s the formula for calculating a correlation from sample data. It’s basically just a way to standardize the covariance between the two variables by dividing it by the product of their standard deviations.

The correlation describes how strong a linear relationship there is between the two variables involved. It ranges from -1 to +1, with zero indicating no relationship at all and +1 or -1 indicating a perfect linear relationship. If you plan to use a correlation to analyze the data, then you need look up the formulas required to do a power analysis for this test statistic in your trusty stats book and take a close look at the relevant effect size measure and power analysis formulas.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Example 2: Test a Correlation

- How much power do I have to detect a correlation of $\rho = 0.3$ if $N = 40$?
  - Given $\alpha = .05$, ES, & $N = 40$ ➔ Expected β
- Measure ES with:

$$d = \frac{2\rho}{\sqrt{(1-\rho^2)}} = \frac{2(0.3)}{\sqrt{(1-0.3^2)}} = 0.629$$

25

This time, we're going to a different kind of power analysis. Let's say you expect to only have about 40 participants available and you want to use power analysis to see how much power you have to detect a correlation of 0.3.

To calculate the expected power, we need to have values for alpha, effect size, and N. Let's stick with the conventional value of alpha = .05 and of course with N = 40. Since we are hoping to detect a correlation of at least 0.3, we use that value for rho to calculate the corresponding effect size. Notice that the effect size measure is still called d here, but that there's a different formula for calculating it now that we're interested in a correlation test instead of a t-test.

The formula requires us to take 2 times the expected correlation and divide that by the square root of the quantity 1 minus the square of the expected correlation. The first thing to notice here is that there's really only one input parameter (rho), though it appears in both the numerator and the denominator. In the numerator increasing rho will cause the numerator to increase, thereby increasing the effect size. Meanwhile, increasing rho in the denominator will cause the denominator to get smaller, which will increase the overall effect size. So, larger expected correlations will always increase the effect size. If we expect a correlation of zero, the whole effect size will be zero.

Plugging in the numbers corresponding to our example assumptions and simplifying the expression gives us an effect size of .629.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Example 2: Test a Correlation

Calculate the noncentrality parameter (δ):

$$\delta = \frac{d\sqrt{N - 2}}{2} = \frac{0.629\sqrt{40 - 2}}{2} = 1.94$$

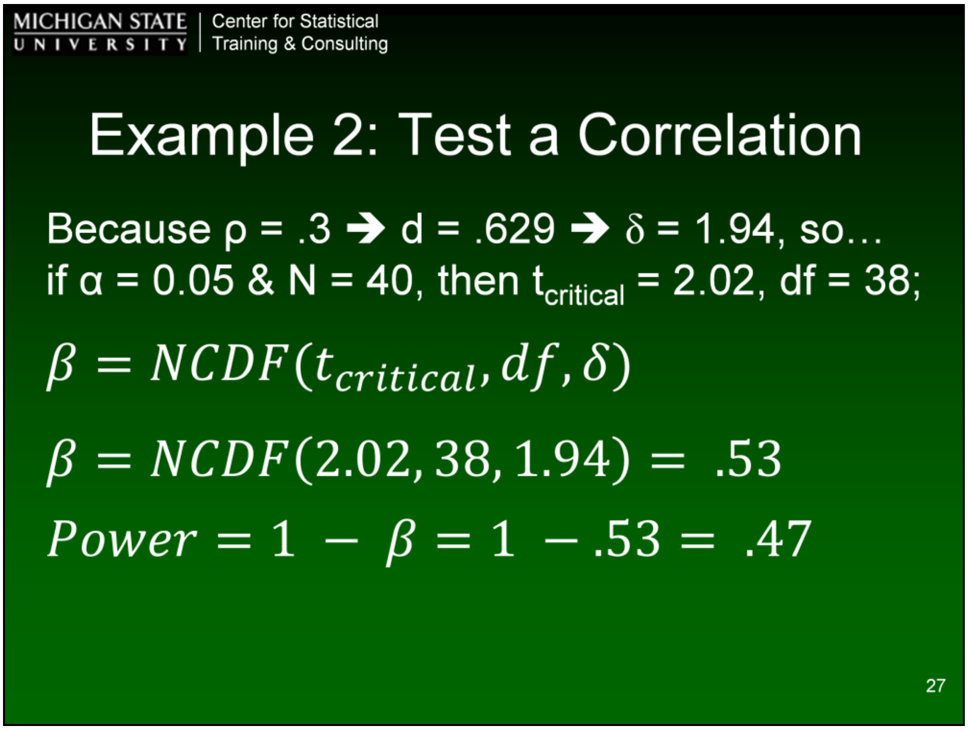
Now we can compute β this way:

$$\beta = NCDF(t_{critical}, df, \delta)$$

26

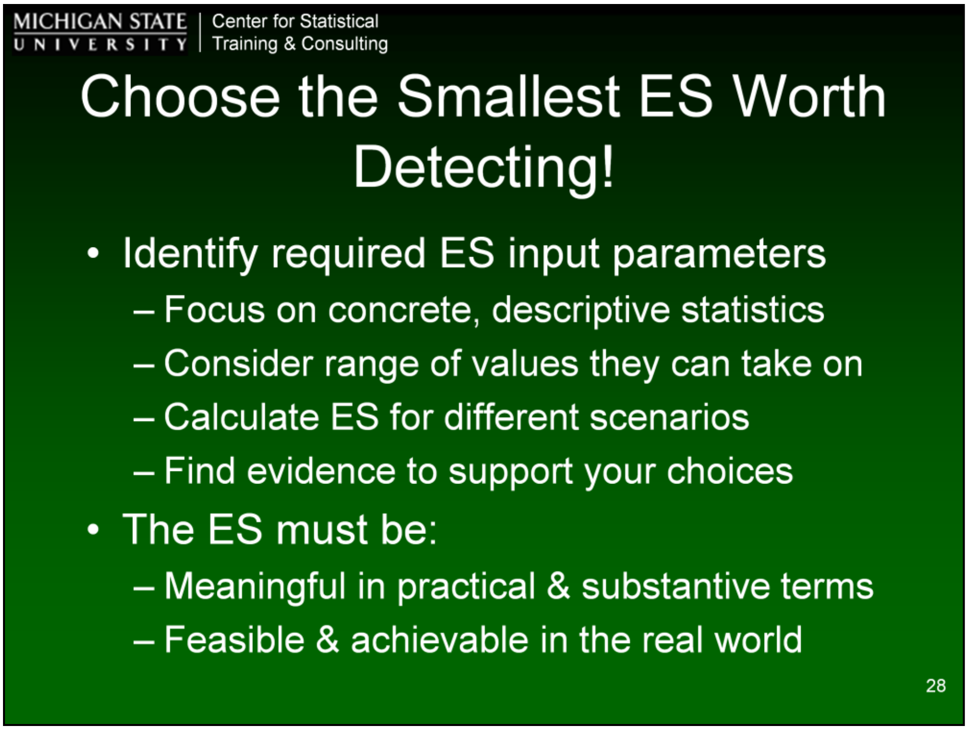
Now that you have an effect size, you can calculate the non-centrality parameter for the correlation test. Notice that this is a slightly different formula for delta than we used for the t-test. That's because this is a different statistical method. However, this time since we are using a form of power analysis where we actually already know both the effect size and the sample size, we're going to use this formula to directly compute the value for delta rather than using the alternate form representing it as the difference between two t-statistics. Then we'll use delta to find the power associated with our scenario. As you can see, plugging in the values for d and N and solving the equation gives us delta = 1.94.

My stats books showed me that we can now use the noncentral cumulative distribution function for the t-statistic to figure out the value of β, which we can easily translate into our final power estimate. In formal terms, β just represents the proportion of the non-central t-distribution that falls below the critical value of t that allows us to reject the null hypothesis at our chosen α error rate of .05. That depends on both the degrees of freedom and on the non-centrality parameter. Because of the complexity of the underlying equations for these types of distribution functions, I recommend using statistical software (or Excel) to calculate things like this.



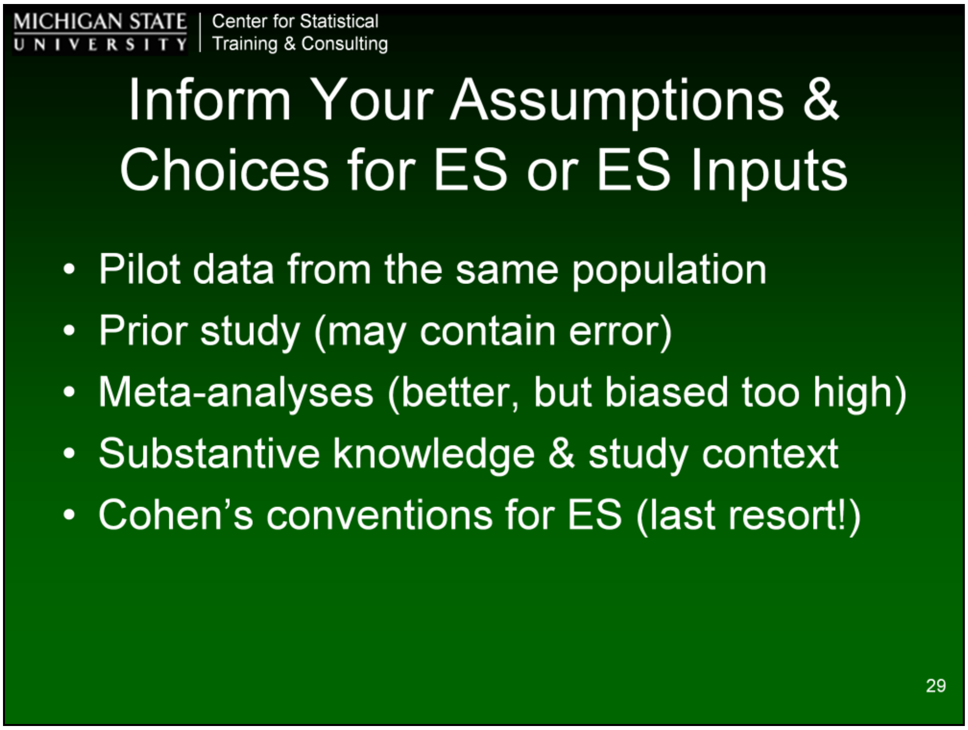
So, let's put it all together and see how much power we would have in this example. Remember we want to detect a correlation of at least .3, so the effect size here is .629. That means our noncentrality parameter delta = 1.94. Given a significance level of .05 and a sample size of 40, the critical value for t at 38 df is 2.02. Once we plug these required numbers into some software, we get a value of beta = .53.

Now, we just subtract that from 1 to get the power for this hypothetical study. We end up with just 47% power, indicating that we will detect a real correlation of .3 less than half the time with this sample size.



Other than choosing an appropriate method, the ES you use in your power analysis affects the answers you get more than any other decision you make, so the numbers you use as inputs to come up with your ES estimate are crucial. Your primary goal should be to identify the smallest ES that is worth detecting. But, how do you come up with that? Start by closely examining the relevant ES formula to see what kinds of inputs it requires. Usually, they will be concrete, descriptive statistics that you can think about and interpret more easily than the final ES estimate. Think about what ranges of values these inputs can realistically take on, then create a few possible scenarios and translate your raw inputs for each scenario into an ES estimate. Then, go find evidence to support the assumptions you'll make about the values to use for each input. I'll come back to this point on the next slide.

Overall, you want to focus on identifying an ES that is meaningful in practical, substantive terms. If your stakeholders would say that the effect you're describing with an ES estimate is trivial and has no real implications, then it's too small and you need to use a larger ES. Say you're looking at the effect of a training program on the subsequent salaries of the physicians who participated in it. Given typical physician salaries, a difference that amounts to $10/year is trivially small. It would take huge amounts of data to detect it, but unless the difference is much larger than that, it has no real-world relevance. Maybe a difference of $1,000/year would be deemed more substantively important. On the flip side, if the effect you're talking about is so large that it cannot realistically be achieved, then you need to accept the fact that you should be looking for a smaller effect size. For example, maybe a delinquency prevention program is highly unlikely to reduce recidivism rates from 60% to 5%, but a reduction to 40% is both achievable and would still have meaningful social impact.



You can use several sources to inform your choices about values for the overall ES or for specific input parameters used to calculate it. Certainly pilot data from the population you intend to study is incredibly useful. It's especially useful to combine what you learn from pilot data with what you learn from these other sources as well.

ES estimates from previous studies of the same phenomenon (or even similar phenomena, alternative programs targeting the same outcomes, etc.) may be useful, but you have to remember that an estimate from any one study is a sample value that may not be good estimate: it will always contain some amount of error (though won't know how much). Furthermore, the authors of previous studies may not have put any effort at all into assessing or explaining to what degree the effects they report are meaningful in practical terms.

Pooled ES estimates from meta-analyses are more likely to be accurate than those from single studies, but even they are likely to be biased toward being too large because studies with non-significant effects often don't get published at all. Use the lower end estimates from meta-analyses to guard against that possible bias.

You always want to use as much subject matter knowledge and information about the study context as you can to inform your choices. Perhaps you can identify what size effect would be necessary to justify the cost of the program.

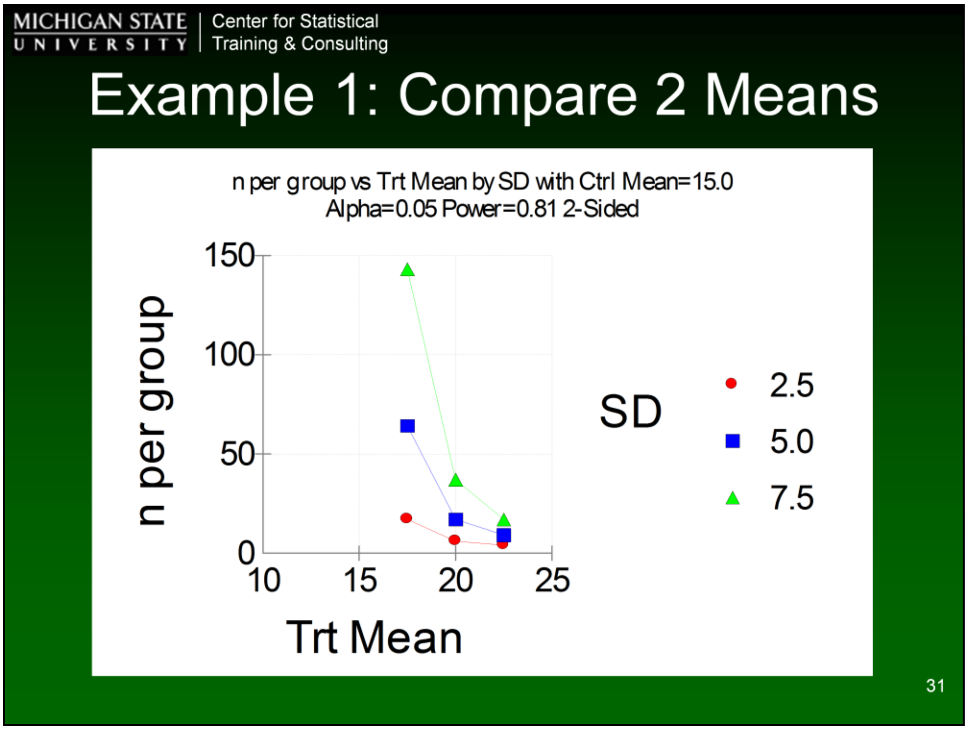
Finally, many of you have heard of Cohen's conventions for what are considered small, medium, and large effect sizes. Using those should be your last resort because they are arbitrary and are not informed by the substantive nature of what you're studying. Cohen didn't intend for them to be so heavily used without any thought about the real nature of what they imply about the specific outcomes in a study.

MICHIGAN STATE UNIVERSITY | Center for Statistical Training & Consulting

Sensitivity Analysis

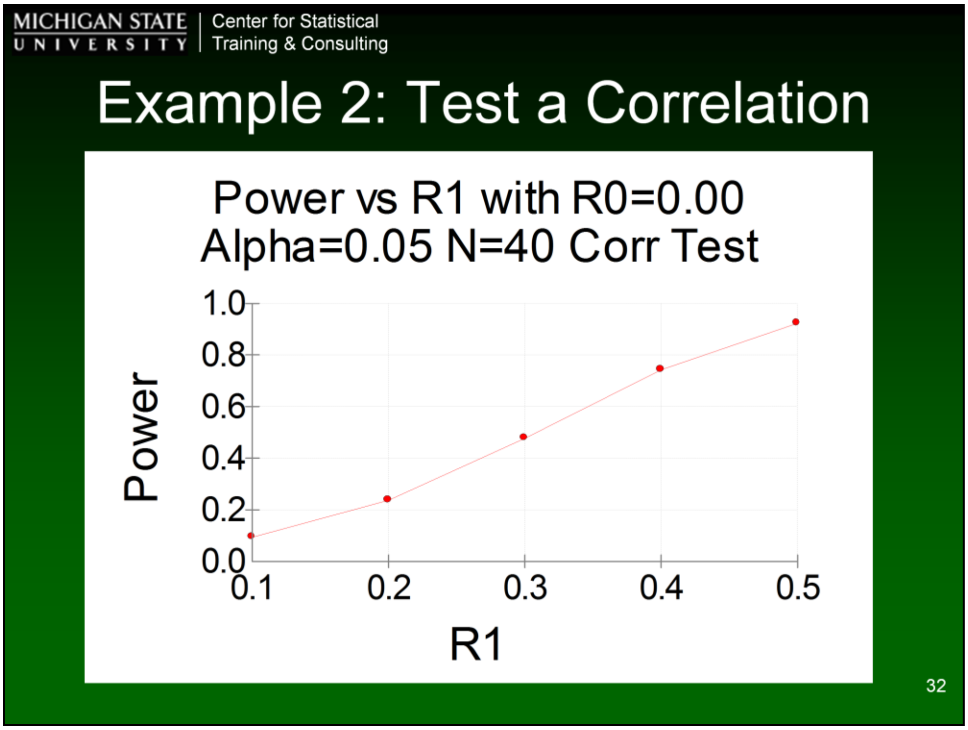
- You have to make assumptions, but…
  - They might be wrong, and…
  - Results may be sensitive to small changes
- Examine the impact of changing them!
  - Try increasing or decreasing ES or N
  - Look at the tradeoffs carefully
- Let's re-visit examples 1 & 2

30



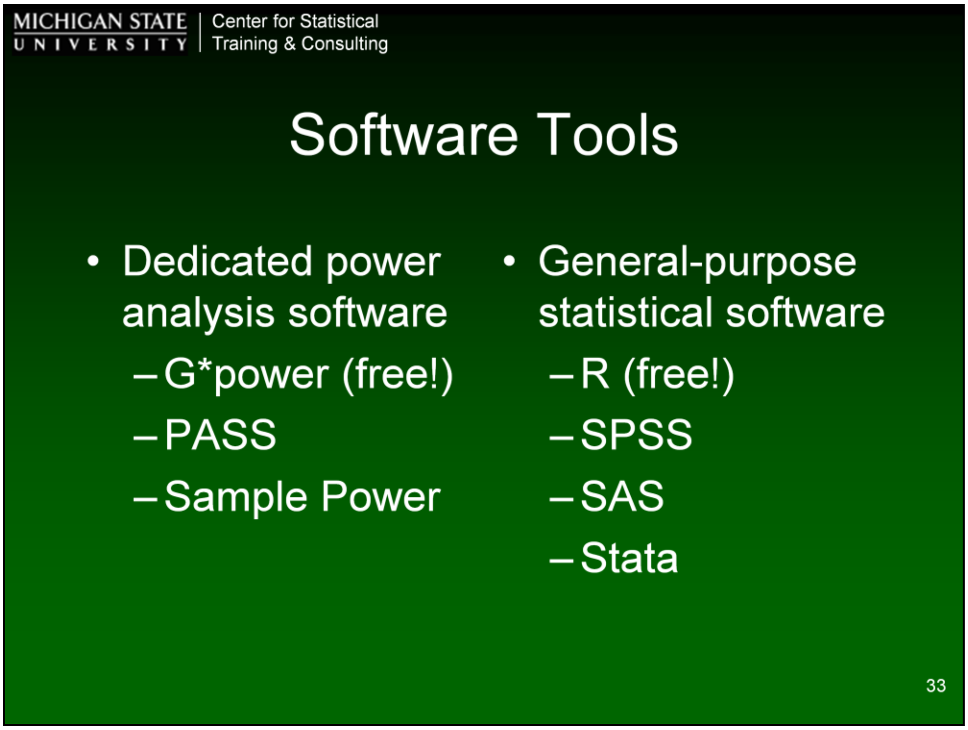
Here, I've used to the PASS 2008 software to do a power analysis for the independent t-test in example 1, while systematically varying the expected mean for the treatment group to be 2.5, 5.0, or 7.5 points higher than the control, group mean. At the same time, I tried three different values for the shared standard deviation (2.5, 5.0, and 7.5 points), creating a total of 9 scenarios. Throughout, I kept the control group mean set at 15, alpha = .05, and desired power set to at least 80%. These scenarios have effect sizes ranging from 0.3 (in the top left point on the green line) to 3.0 bottom right point on the red line. The required number of people per group varies tremendously across these scenarios.

Power	Required n per group	Ctrl Mean	Trt Mean	SD	d
81%	17	15.0	17.5	2.5	1.0
80%	64	15.0	17.5	5.0	0.5
80%	143	15.0	17.5	7.5	0.3
88%	6	15.0	20.0	2.5	2.0
81%	17	15.0	20.0	5.0	1.0
81%	37	15.0	20.0	7.5	0.7
94%	4	15.0	22.5	2.5	3.0
85%	9	15.0	22.5	5.0	1.5
81%	17	15.0	22.5	7.5	1.0



Here I've used the PASS 2008 software to graph the amount of power you have versus the size of the correlation used to generate the effect size estimate. Previously, we found that if your sample size is 40 and alpha = .05, then you have power = .47 to detect a correlation of .3. Here we can see what happens to power if the correlation is actually as low as .10 (where power is a mere 9%) or as high as .5, where power is 92%.

I wish I had time to present more examples, but I really wanted to share the basic concepts in a way that would help you see both why power analysis is important and why it's so important to think carefully about the assumptions and decisions you have to make along the way. If you're working with a statistician, you'll need to help him or her find values for some of the input parameters used in the power analysis.



Software tools are usually either dedicated power analysis software, or general purpose statistical software that has features you can use to do power analyses. I generally recommend using dedicated power analysis software because it tends to be more user-friendly because you get clean, simple menus and screens for selecting procedures. These often have helpful hints and information built right into the user-interface or easily accessible via the help system. G*Power is a fairly decent piece of free software that you can download from the web. It supports a fair number of different statistical tests, but compared to some of the commercial alternatives, you have to know more about power analysis to use it effectively. However, you can get even better software if you're willing to pay for it.

I've been using PASS 2008 pretty extensively. Although it costs several hundred dollars, it is very comprehensive (it covers a lot of different statistical methods), plus it has a fantastic user-interface and help system that gives you lots of information about how to set various parameters. The output includes citations for the books and papers describing the computational methods it uses for each procedure. It's great for doing sensitivity analyses to see what happens when you vary the parameters. Version 11 of PASS was just released a week or so ago, and it looks like a nice improvement over the previous version that I've been using.

Of course, most of the general statistical packages on the market can also do power analyses. R is free, open source statistical software. It's incredibly flexible and powerful (it's actually a programming language dedicated to doing statistics), but it is not very user-friendly to people who are not comfortable with computer programming. There are some user-contributed modules you can download for R that add specialized functions for power analyses, but you have to know how to find them and it still demands writing short programs.

Chris Aberson's recent book on power analysis shows how to use SPSS to do a fair selection of power analyses. The downside of using the general statistical packages is usually that they aren't as user-friendly as specialized power analysis software like PASS or Sample Power, so you often have to be more knowledgeable and skilled to use them for this purpose.

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences. New York, NY: Routledge.*


MICHIGAN STATE UNIVERSITY
Center for Statistical Training & Consulting
Slides & suggested resources will be available soon:
• My website: www.msu.edu/~pierces1
• AEA public eLibrary: http://comm.eval.org/
• Via e-mail: pierces1@msu.edu
Time for questions & discussion!
34