Statistical Power: How Not to Miss What’s Right in Front of You

Erin M. Buchanan1

1 Harrisburg University of Science and Technology

Author note

Special thanks to Twitter and Joe Simons for suggesting great titles for this chapter.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA 17101. E-mail: [ebuchanan@harrisburgu.edu](mailto:ebuchanan@harrisburgu.edu)

Abstract

In this chapter, we discuss the definitions of power and how to interpret power in Null Hypothesis Significance Testing and how these interpretations can be extended to Bayesian frameworks. Next, the main determinants of power are outlined including the sample size, effect size (and variability), , and the type of statistical test. Each influence on power is demonstrated with example studies on statistics education and data literacy. Different types of power analyses, planning for sample sizes and sensitivity, are illustrated using traditional charts, popular programs, simulation, and accuracy in parameter estimation. Last, the limitations of power—especially what it does not tell you and what you should not do—are outlined to warn you about the potential misuses of power analyses. Suggestions on appropriate power planning are provided at the end of the chapter.

Statistical Power: How Not to Miss What’s Right in Front of You

Statistical power is one of the most difficult aspects of designing and implementing a research project, as it often is confusing and overly technical to many scientists. One reason for this confusion is the number of facets that serve a role in understanding or calculating power, that all need to be considered simultaneously for the best outcomes in a research project. First, let’s cover the definition of power and related statistical concepts, so you can understand how to use power analyses to plan your research studies. Three example studies related to statistics education and literacy will help solidify how to apply the concepts of power and calculate power analyses for future studies.

# What is Power

## NHST Framework

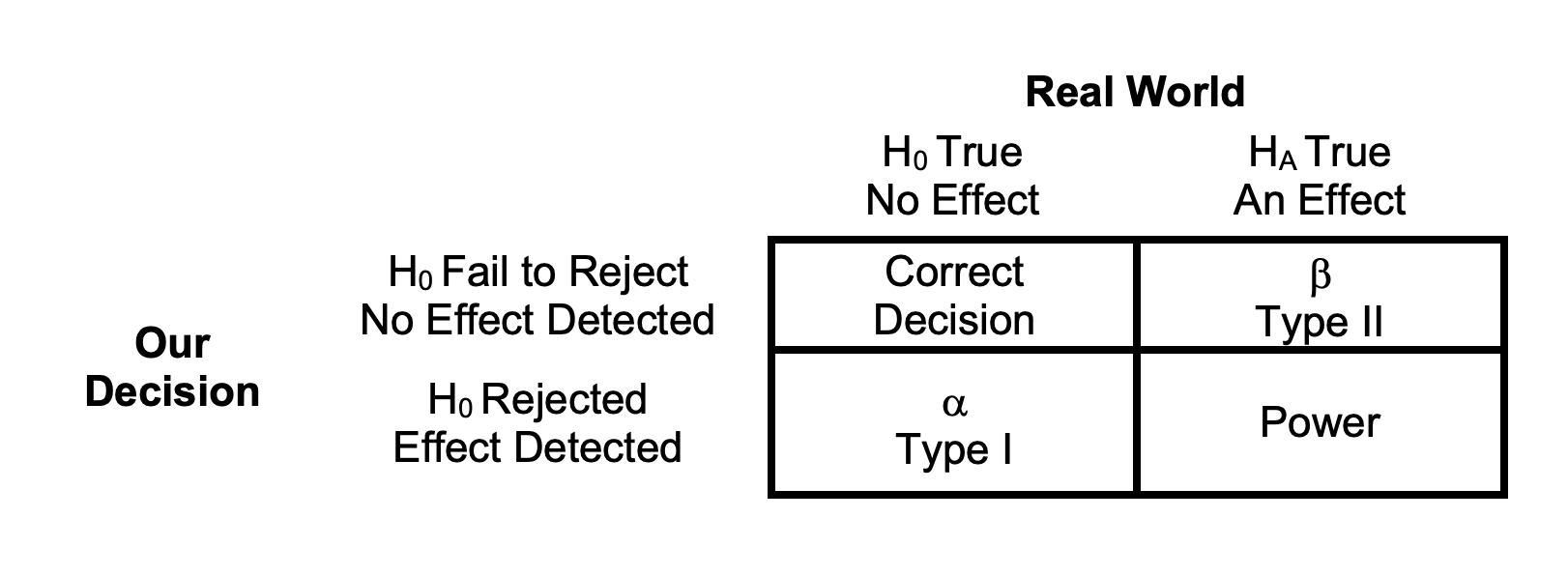
When using Hypothesis Significance Testing (NHST), statistical power represents the ability to discover an effect *if* it truly exists. NHST, in the modern era, is a combination of the works from Neyman-Pearson and Fisher used to test opposing hypotheses (Dienes, 2008). The null (or nil) hypothesis is generally the default consideration that no effect/relationship exists. The alternative (or research) hypothesis suggests that something did happen in the study. For example, in one study (outlined below), we might propose:

Null Hypothesis: The classes taught with *R* and *JASP* do not have different final exam scores.

Research Hypothesis: The classes taught with *R* and *JASP* have different final exam scores.

In this set up, the hypothesis does not predict a specific direction (i.e., *R* students performed better than *JASP* students), but rather simply predicted that some difference between their final exams may be found. We could use an independent *t*-test as our statistical test to determine if we find any differences between classroom final exam means. Statistical tests can be imagined as a ratio of our model hypothesis to the amount of error in the study. For example, we want to detect if there’s a larger difference in the final exams between classes (model hypothesis) than we might expect considering individual students are all different (error). Power is the ability to find a high enough ratio to detect the difference.

Within the NHST framework, the researcher must decide to either “reject” the null hypothesis as very unlikely given a large enough ratio of model to error or “fail to reject” the null hypothesis when the ratio of model to error is small or equal. In Figure 1, we show our decision on the left-hand side, and on the top, the real world, if one could know the real outcome, is displayed. If we failed to reject the null hypothesis, we would make the correct decision if there was no real difference in final exams. However, if we fail to reject the null hypothesis (do not detect those differences), and there truly are differences in R and JASP classrooms, we have “missed” detecting the effect (i.e., or a Type II error; top right corner). On the other side, if we reject the null hypothesis, and say there are differences in final exam scores, we could make a Type I error (represented by ) if there was no actual difference in reality. Last, power reflects the probability that we would reject the null hypothesis when we should, as there is a difference to be found between final exam scores.



*Figure 1*. Decision errors, power, and correct rejections for null hypothesis testing.

Let’s provide an analogy to help solidify this type of thinking. Imagine you are a detective who is trying to solve a crime. Your statistical test is the amount of evidence that you can collect that points to who committed the crime (model) compared to the evidence that does not clearly support that they did the crime, or the evidence is wrong (error). A Type I error would occur if person A was found guilty of the crime, as it appeared there was enough evidence, but in reality, person A did not commit the crime. A Type II error would occur if person A committed the crime, but you could never find enough evidence to convict them. Last, power would be the ability of the detective to find the evidence to convict the right person for the crime. Another famous example is of pregnancy tests: a Type I error would occur if a someone who is not pregnant was told they were pregnant at the doctor’s office, while a Type II error would occur if the pregnancy test was negative and did not detect an actual pregnancy.

Our ability to know the “real” answer is generally limited, especially in scientific studies. Therefore, α, β, and power are generally represented as theoretical concepts used to control and plan research studies. Each of these values is expressed as a probability. Researchers will generally set the probability of α to ≤ .05 to avoid a Type I error, as it is perceived as more problematic than a Type II error and is a strong research norm. Several fields , such as medicine and physics, require lower α values for their studies, usually α to ≤ .001. β is commonly set to ≤ .20—a 20% probability of missing a true effect. Power is the direct opposite of β, and therefore, by choosing .20, power becomes .80—an 80% likelihood of detecting an effect that exists (i.e., power = 1 - β). Note that the probabilities listed here are common practice in current research but are not necessarily “correct”, and each should be carefully considered and justified within the context of any research study (Lakens et al., 2018). Within the current credibility revolution—the drive for improved quality in scientific studies (Vazire, 2018)—power plays an important role, and many have suggested setting this value to a higher standard (i.e., power > 90%) to improve the evidence provided within an individual study.

The rest of this chapter focuses on power analyses as defined for NHST frameworks. Other statistical frameworks, such as Bayesian analysis, still require power analyses (Dienes, n.d.; *Statistical Rethinking*, n.d.). However, they may not use traditional or popular programs that others use for power purposes that we will outline below. Readers interested in learning more about Bayesian analyses can check out Dienes (n.d.).

# Study Examples

To help illustrate these issues, we will consider three different research studies focusing on statistics education that highlight the nuance required for statistical power analysis.



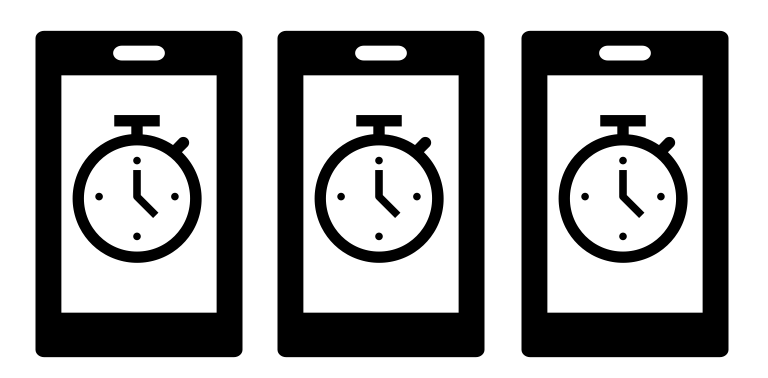
*Figure 2*. Study 1 illustrates two separate statistics classrooms each learning a different software to determine the influence of computer programs on final grades.

In our first research project shown in Figure 2, a statistics teacher is interested to know if students differ in their final exams scores based on the software used within an introduction to statistics course. Several professors have implemented the software *R*—a programming-based computer language that was designed to calculate statistics (R Core Team, 2022). Learning a programming language is often daunting to students, and the statistics teacher also recruits several instructors who use *JASP* (Jeffreys’s Amazing Statistics Program)—a point and click statistics program designed to calculate statistics using *R* with no programming knowledge (JASP Team, 2022). Both programs are desirable to students, as they are both free, unlike their rivals.



*Figure 3.* Study 2 illustrates a survey design where student’s data literacy is assessed to understand its relationship with future employment in data analytics.

Our second study in Figure 3 illustrates a survey in which student’s data literacy and digital competence is measured at the end of their college career in a university exit exam (Gümüş & Kukul, 2023). Digital competence refers to the ability of individuals to process, understand, and apply information found in digital sources (e.g., social media, news outlets, research, Pangrazio et al., 2020). Not only are these skills necessary for day-to-day activities (i.e., interpreting the weather report), they are increasingly necessary for employment. Trends in job descriptions and hiring advertisements indicate that employers are increasingly interested in data analytics and digital competence (*Data Scientists* , n.d.; Matli & Ngoepe, 2020). Data analytics can be leveraged across the workforce from research careers in analyzing study data, in education to improve student learning, and business practices to optimize performance, increase profits, and improve customer relations. Software tool kits, such as *R* and *JASP* are featured in data analytics as well as the ability to comprehend, process, and apply statistical skills. In this study, the correlation is examined between students’ data literacy scores and self-reported income in their first job after graduation.



*Figure 4*. A depiction of Study 3 that examines the implementation of gamification into a statistics classroom to improve data literacy.

Finally, our last study extends into a newer methodology: ecological momentary assessment [EMA; Shiffman et al. (2008)]—shown in Figure 4. In EMA studies, participants are asked to fill out smaller surveys or questionnaires over a period of time (e.g., once a day for two weeks, once a week for a semester). This type of design has surged in popularity, as survey platforms and special smartphone apps have made it easier for both the researcher and participant to complete the study (Porras-Segovia et al., 2020). As noted, data literacy is increasingly imperative for future employment, yet many students have anxiety when taking a statistics course (Aerts et al., 2021; Ralston, 2020). In this study, statistics courses were “gamified” by adding challenges and leader boards to the course to mimic game play found in many popular phone and video games (Legaki et al., 2020). Students were encouraged to maximize their points and compete in teams to improve on the leader board, and these game modules served as practice to improve data literacy skills and improve student performance on exams (Seaborn & Fels, 2015). Students were assessed on data literacy throughout the semester using the EMA design, and these scores were predicted by the game module points in the course.

In this chapter, we will use these three studies as examples to outline the definitions and influences on power. In the next section, we will explore what factors can influence power. These influences will be applied when demonstrating how to calculate power and how the same design can lead to different answers when exploring power. Last, we will warn you about what a power analysis *does not* tell you as a cautionary tale to understand the limitations of power.

# Determinants of Power

In this section, we will review the factors that influence power including sample size, the effect size, α, and the type of statistical test. For each, you will see an example of the factor’s effect on overall power. Remember, while we discuss these individually, they tend to interact within a study—as each factor changes, the power will change accordingly. That said, researchers tend to fix most of these values and focus on one influencing factor at a time, when conducting a power analysis, to decrease complexity of calculation.

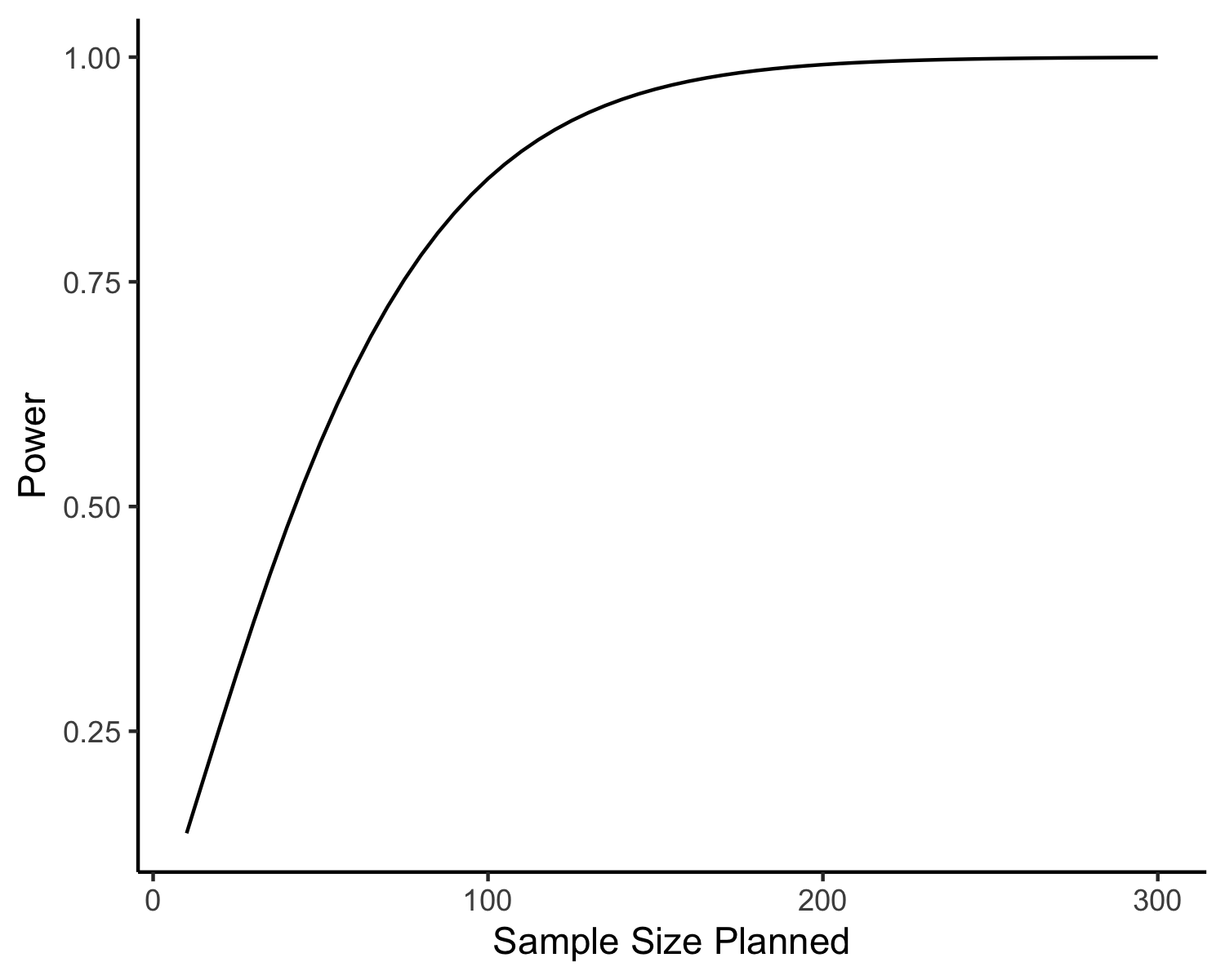
## Sample Size

Sample size is often considered the main determinant of power in a study, especially because sample size is the main factor that a researcher has control over when planning and executing a study. As such, it is typical to see a published or pre-registered “power analysis” focus on determining sample size necessary for their study.

So why does sample size have such a large influence on power? Remember that statistical tests are generally calculated by using a formula of model variability to the amount of error variability. Error variability is generally some form of standard error and is calculated by or a similar form depending on test. This formula represents the standard deviation—the amount of variation around the mean—divided by the square root of the sample size (*N*). With this formula, as *N* increases, the overall standard error decreases. When you divide your model variability by a smaller error variability (standard error), the ratio of model to error increases, and the ability to find a significant effect also increases. In fact, this influence is so well known, it is a common statistical joke to note that with large enough sample sizes “everything is significant”. As we will note in our limitations section, large power and significant results are not necessarily practically important results.

Large sample sizes have additional bonuses beyond increasing power. First, we have increased our precision in measurement (see the Newer Methods - AIPE section below) for our study and hopefully sampled a larger diversity of participants. Larger samples mean that individual data points have less influence on statistical parameters, such as the mean and standard deviation. Last, the “big-team” science movement has made collecting large sample sizes easier on individual researchers. In these studies, multiple research labs, often across the globe, join to collect data (Coles et al., 2022; Cuccolo et al., 2021; Koch & Jones, 2016) to improve power, generalizability, and scientific knowledge.

In our second study, we will examine the correlation between data literacy and first job income. If we set all the other variables that affect power, how does the sample size increase our ability to detect a correlation? As you can see in Figure 5, larger sample sizes increase power in a curvilinear fashion (often called a power curve!) that eventually approaches nearly 100% power.



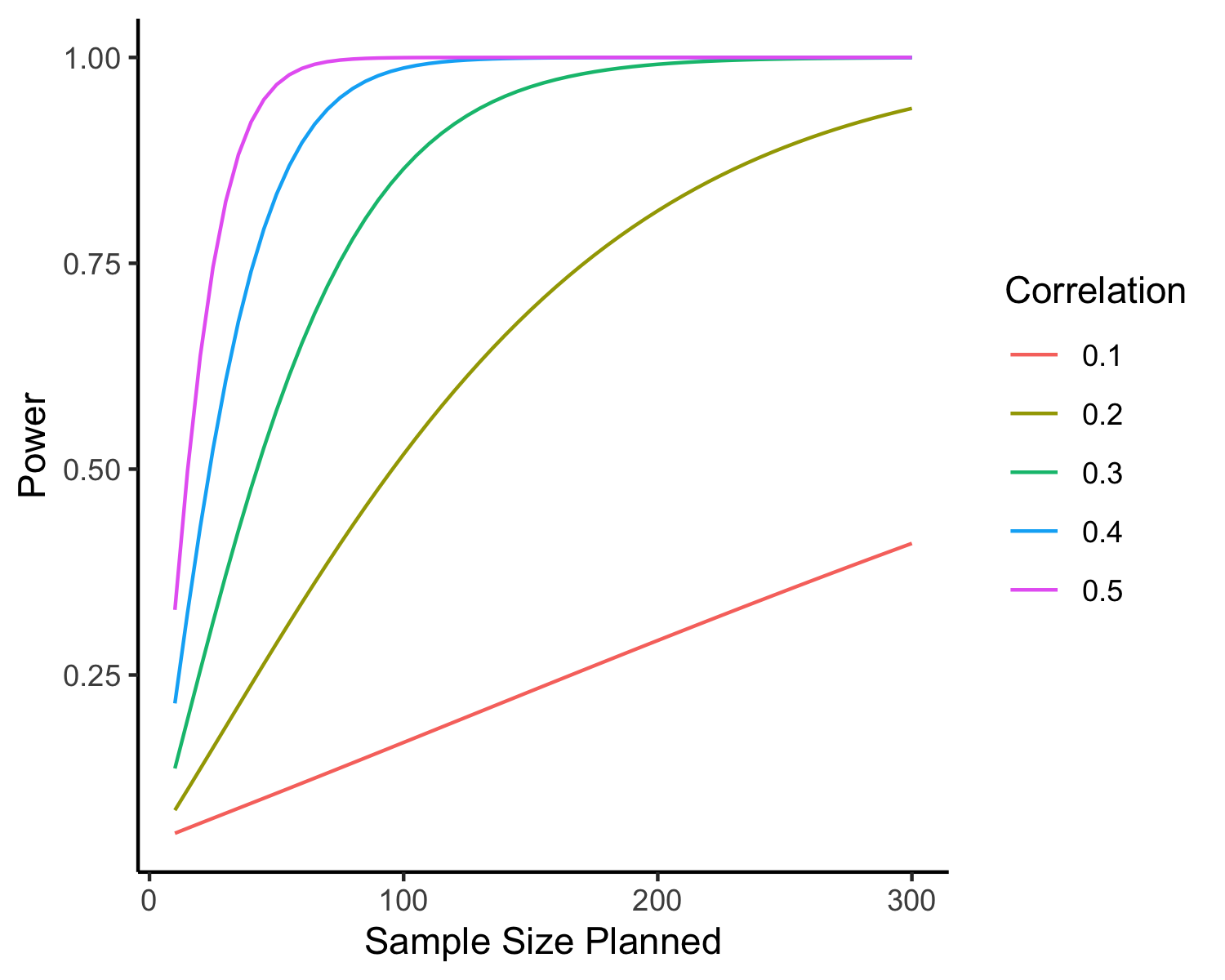
*Figure 5.* The relationship between planned sample size and power in the correlation study measuring data literacy skills and future income. This curve represents the power and sample size for a correlation of *r* = .30 using of .05.

## Effect Size

Effect size is our second factor that influences power. Effect size is a standardized measurement of the magnitude of the results in our study—–that is, how “big” the results were in the study (Lakens, 2013). Some common effect sizes include *d* (the standardized mean difference between groups or measurements), *r* (the correlation between two measurements), and (the proportion of overlap between the independent variable(s) and the dependent variable). Effect sizes can aid researchers in comparing between studies with different variables, interpreting the size or impact of their study variables, and planning future studies.

Effect sizes are generally impacted by two components: 1) the size of the model variability, and the 2) size of the error variability. As noted earlier, these two pieces are the formula for most statistical tests and part of our explanation for why sample size is a large factor in power analyses. The difference between the statistical test formula and effect size formula is the sample size: if the statistical test is approximated by then the effect size would be , as effect sizes are designed to represent the magnitude of the model without the influence of sample size. Note that this explanation gives you the basic idea of the concept for effect sizes, as the real formulas for effect size are influenced by the number of variables and type of research design.

Power will increase with larger effect sizes, as it’s always easier to find effects that are larger. As the size of the model variability increases, we see increases in power; correspondingly, as the model error decreases, we also see increases in power (because it’s in the denominator of the formula). If we add a few different effect sizes to our data literacy study (rather than just *r* = .30), we can see the impact of effect size and sample size simultaneously in Figure 6. As the different correlations increase in size (the lines on the figure), the power increases at each sample size. We can see how they interact, as larger effect sizes show a sharper increase in power as sample size increases because the lines reach nearly 100% power faster (e.g., they are steeper).



*Figure 6.* The relationship between sample size planned, power, and effect size (the lines) for our correlation study measuring data literacy and income using = .05.

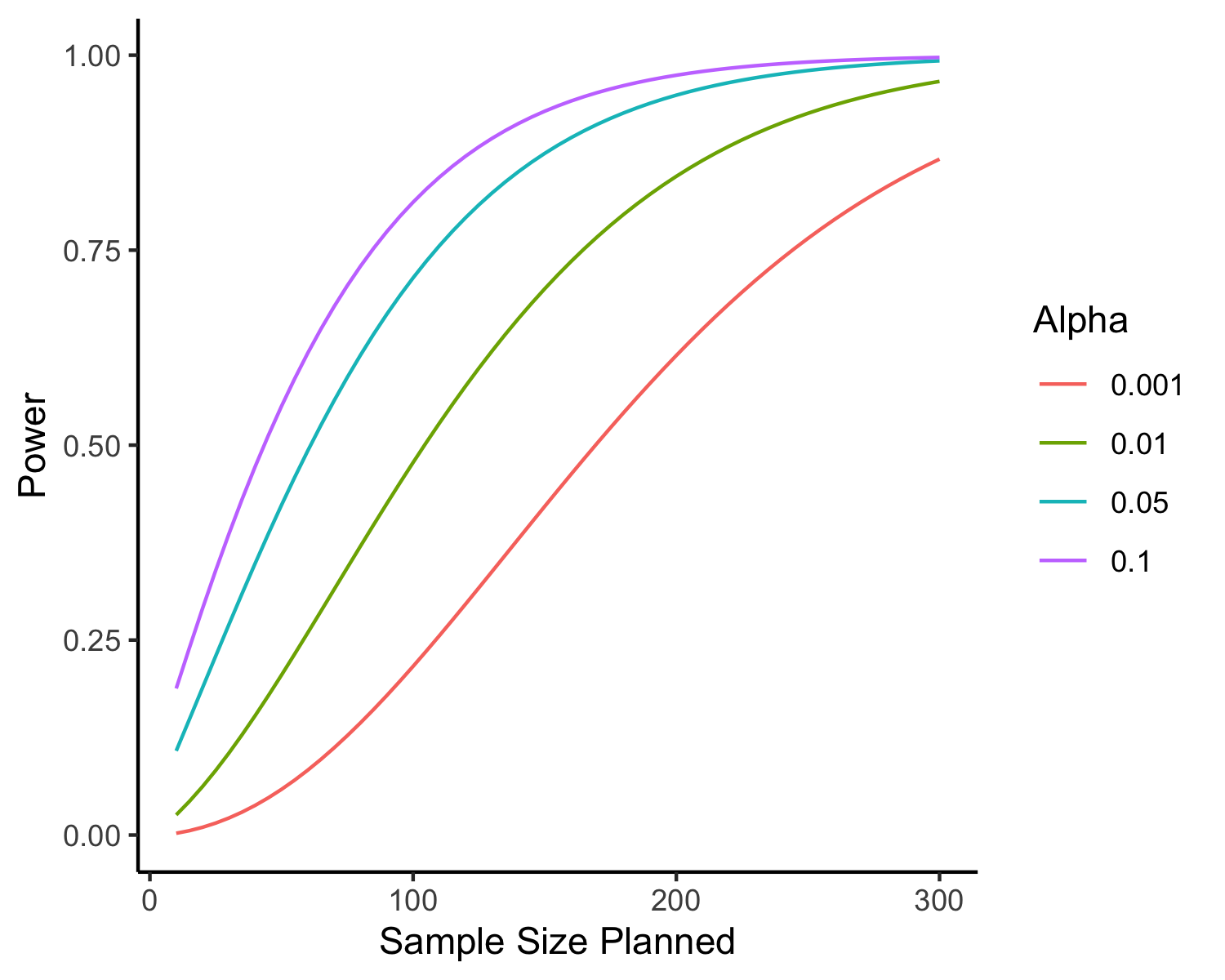
## Sample Size and Effect Size Together

We noted in the previous section that many power analyses are used for study planning to determine the sample size necessary for the study to achieve a certain level of power. Alternatively, if our sample size is limited due to money, time, or a small study population, we can use a sensitivity analysis to determine the smallest detectable effect given all other considerations (Vevea & Woods, 2005). These analyses have become more popular to illuminate if the study has or had adequate power to detect a certain effect, especially if the study did not find the effect size researchers planned. For example, if the study was planned with a medium effect but found a small one, did they have enough power to detect the small or medium effect? These sensitivity analyses may be especially important when considering publication bias—–the issue that studies are significantly more likely to be published if they contain significant effects (Franco et al., 2014). If we use previously published results to help estimate our effect size for a newly planned study’s power analysis, it is likely that we will use a number nearly twice the real effect size (Open Science Collaboration, 2015) because the studies that did not show larger effects did not get published. Both sensitivity analyses and smallest detectable effects will be demonstrated below in power analysis section.

## Alpha

As described earlier, α is the Type I error rate (i.e., the probability of incorrectly rejecting the null hypothesis when the null is true). The most common α value found in social and behavioral science literature is α < .05, although there is no particular reason why it should be set to .05. Some scientists recently suggested this value should be lowered to control for the number of studies that were published with false positives (Benjamin et al., 2018), while others suggested that the rule was not the problem—–the mindless use of the rule was the issue (Lakens et al., 2018). Of note, physics and the medical field often use lower criterions of .01 or .001 (Meehl, 1967), so alternative alpha levels are not unheard of. Each individual researcher should examine what they would consider as their threshold before the research is conducted based on their desire to balance Type I to Type II errors (Miller & Ulrich, 2019).

As increases (e.g., as it approaches one), power also increases. Using our sample size example from above, we can plot the power for several values in comparison with power (Figure 7). The interpretation here is that if you give yourself a 10% chance of falsely rejecting the null ( = .10), you will increase power versus only giving yourself a .1% chance of falsely rejecting the null ( = .001). It is usually not recommended to simply increase α to increase power, especially as the scientific community worries that too many false positives already exist.



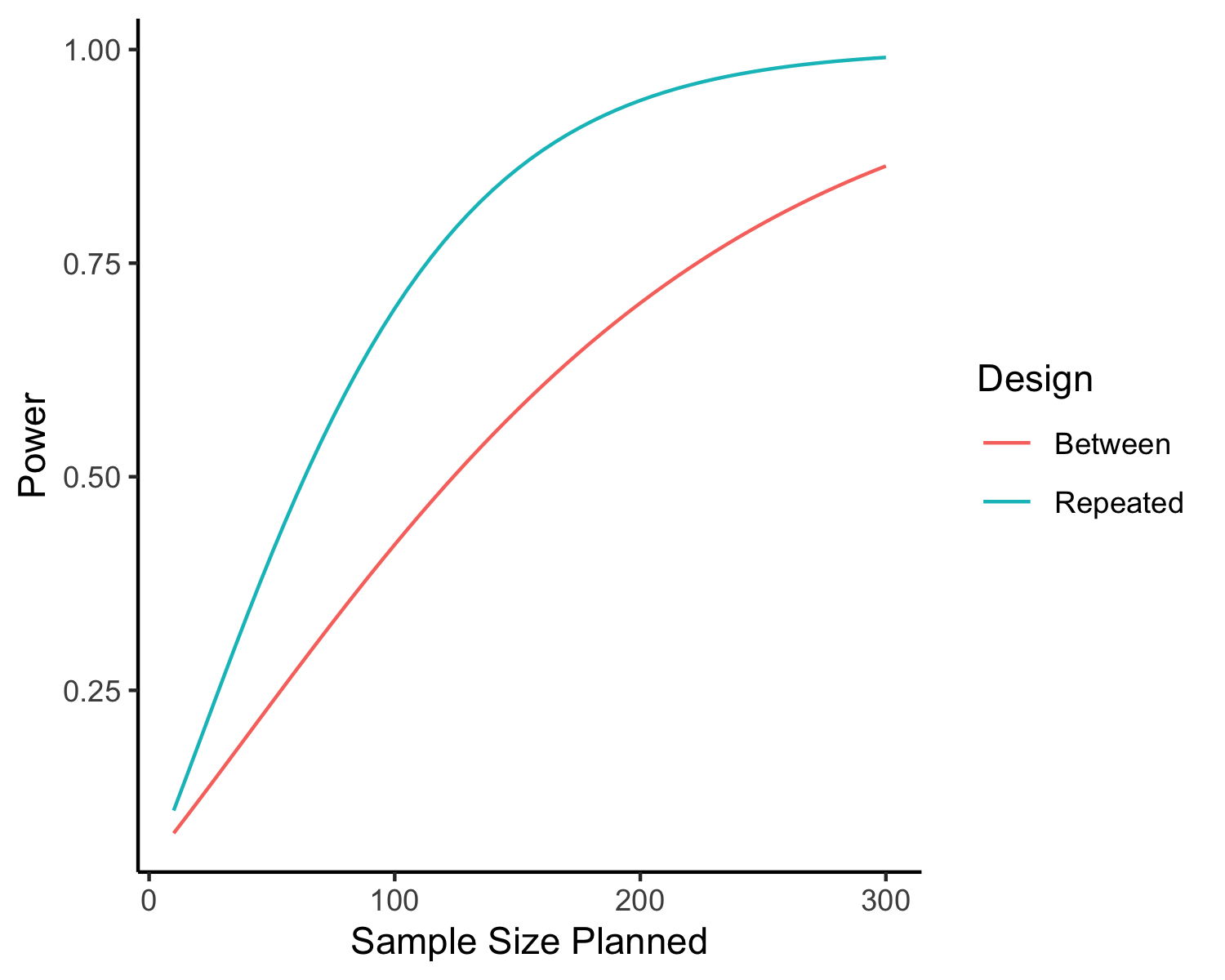
*Figure 7.* The relationship between sample size planned, power, and the criterion for our data literacy correlation study where *r* ~ .25.

## Test Type

The last main influence on power is the type of research design and statistical test. Although many statistics books pretend that a specific research design is always analyzed with a corresponding specific statistical test, reality is much more complicated. A research study called the “Many Analysts” paper gave multiple teams of researchers the same data and hypothesis and asked each to analyze the data. The study investigators then compiled the number of ways the data was analyzed, and the effect size found in each team analysis. These teams came up with over twenty ways to analyze the data with a wide range of final effect sizes reported (Silberzahn et al., 2018). Therefore, we see that different types of statistical tests can be performed on the same data. The main idea is that some statistic tests are more sensitive, meaning they are likely to give you more power to detect effects that exist in comparison to other statistical tests.

The influence of research design is a bit easier to conceptualize: repeated measures designs have more power than between subjects designs. Please note that these often have multiple names—repeated measures can be called dependent or within subject designs; between subjects can be called independent designs. Our third study is a repeated measures design because each person is measured on their data literacy multiple times within the same study; the first study is a between subjects design as each person is only measured in one classroom for their final exams. Repeated designs have more power for similar reasons as larger sample size and effect sizes have more power—the amount of error variability is *generally* smaller in repeated designs. When people are measured multiple times, we have to mathematically control for their non-independence (e.g., that each response is dependent on a previous response) and controlling for their individual variability tends to reduce the error variability in the model to error ratio.

To help you understand why this happens, imagine you know that some students are great students, and some students are average students. You can control for their individual differences and just look at the change in their scores across time. If you did not know why some students scored higher than others because you only measured them once, all those differences are considered error in your study. The first study is a between subjects design—you compare two different sets of individuals in two separate classrooms. However, you could convert this design to repeated measures by comparing exam scores from two different semesters on the same people (i.e., they learn *R* in one semester and *JASP* in another semester). As shown in the figure below, the repeated measures design would achieve much higher power for nearly all sample sizes.



*Figure 8*. The relationship between sample size planned, power, and the type of research design if we could measure our study on statistics classrooms in both a between and repeated measures design.

# Power Analyses

## Using Tables

In the days before we all carried tiny computers in our pockets, calculating power or estimating the ideal sample size usually involved a table or chart printed in a large textbook (Cohen, 2013) or rules of thumb such as *N* > 30 for groups and *N* > 100 for correlations. These tables usually included sample size estimates for a range of power and effect size estimates. For example, in Table 1, if you wanted to estimate the number of participants necessary for 80% power with an effect size of *d* = 0.50, you would find the necessary row and column to determine you need *n* = 64 participants. These tables were printed for different levels of α and types of test but were very tedious to search through for a specific combination of requirements.

*Table 1*

A traditional power table outlining effect size, power, and planned sample size for independent *t*-tests.

|  | Effect Size (d) | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Power | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 1 | 1.2 | 1.4 |
| 0.25 | 332 | 84 | 38 | 22 | 14 | 10 | 8 | 6 | 5 | 4 | 3 |
| 0.50 | 769 | 193 | 86 | 49 | 32 | 22 | 17 | 13 | 9 | 7 | 5 |
| 0.60 | 981 | 246 | 110 | 62 | 40 | 28 | 21 | 16 | 11 | 8 | 6 |
| 0.67 | 1,144 | 287 | 128 | 73 | 47 | 33 | 24 | 19 | 12 | 9 | 7 |
| 0.70 | 1,235 | 310 | 138 | 78 | 50 | 35 | 26 | 20 | 13 | 10 | 7 |
| 0.75 | 1,389 | 348 | 155 | 88 | 57 | 40 | 29 | 23 | 15 | 11 | 8 |
| 0.80 | 1,571 | 393 | 175 | 99 | 64 | 45 | 33 | 26 | 17 | 12 | 9 |
| 0.85 | 1,797 | 450 | 201 | 113 | 73 | 51 | 38 | 29 | 19 | 14 | 10 |
| 0.90 | 2,102 | 526 | 234 | 132 | 85 | 59 | 44 | 34 | 22 | 16 | 12 |
| 0.95 | 2,600 | 651 | 290 | 163 | 105 | 73 | 54 | 42 | 37 | 19 | 14 |
| 0.99 | 3,675 | 920 | 409 | 231 | 148 | 103 | 76 | 58 | 38 | 27 | 20 |

## Using Programs (Sample Size)

\*Power (Erdfelder et al., 1996; Faul et al., 2007) \*\* powerandsamplesize.com and <https://designingexperiments.com> (Anderson et al., 2017).

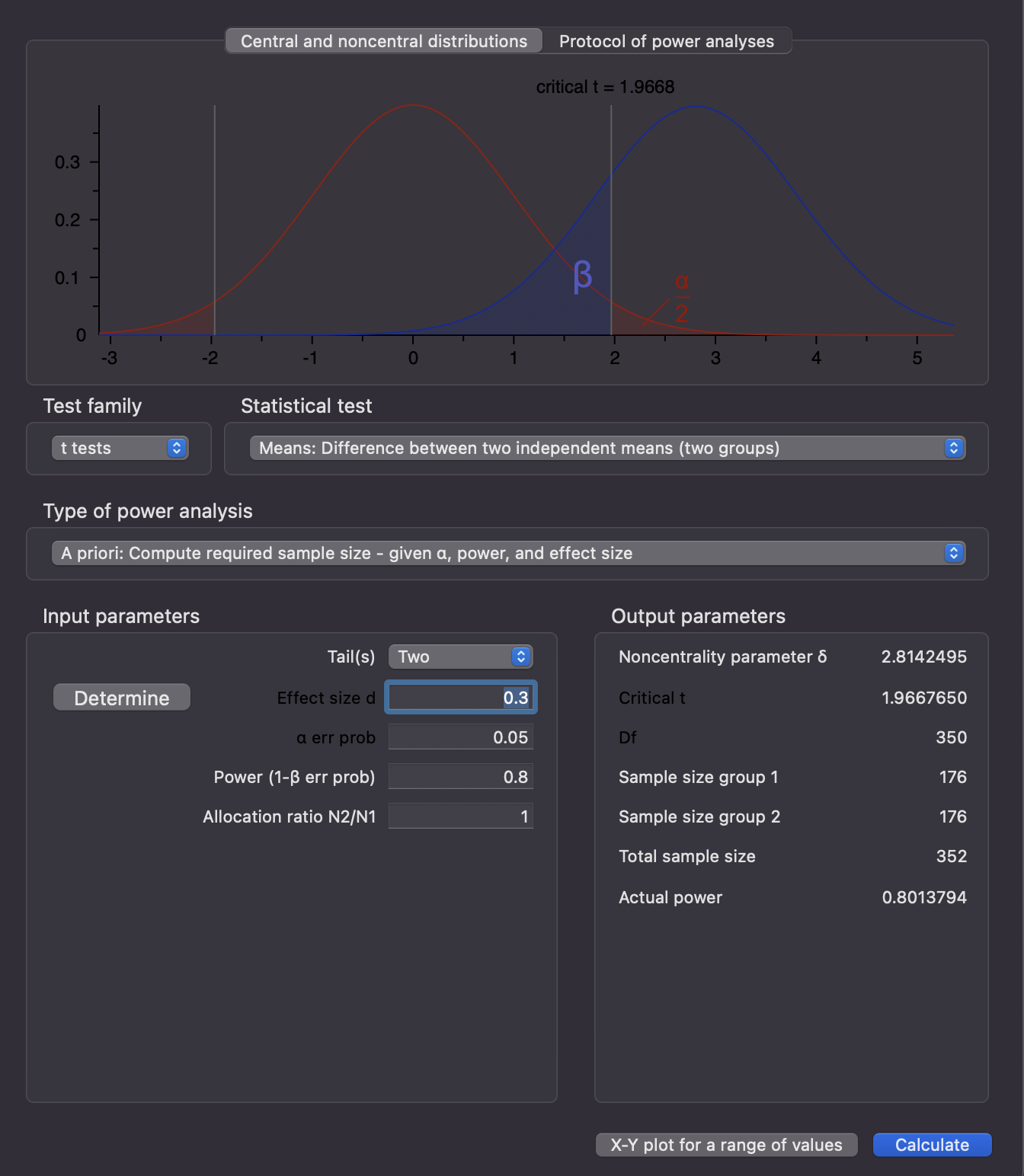
Increases in computing power have improved our ability to estimate and determine information about the power in our studies. Computer programs can use exact formulas to calculate values for sample size, effect size, or precise power values given input by the researcher. The most popular point and click program is G\*Power (Erdfelder et al., 1996; Faul et al., 2007)that is both simple and free. Users like G\*Power because it contains most common statistical tests and has a user-friendly interface for entering the necessary information to calculate a specific value. A table of G\*Power inputs for common statistical tests taught in statistics is provided in Table 2. Web-based tools such as <https://powerandsamplesize.com> and <https://designingexperiments.com> harness the power of computational languages (e.g., *R*) to create online applications for power estimation with simple text entries (Anderson et al., 2017).

*Table 2.*

Where to find common statistical tests in G\*Power menus

| Common Test Name | Test Family | Statistical Test |
| --- | --- | --- |
| Correlation | Exact | Correlation: Bivariate normal model |
| Single sample t-test | t tests | Means: Difference from constant (one sample case) |
| Independent t-test | t tests | Means: Difference between two independent means (two groups) |
| Dependent t-test | t tests | Means: Difference between two dependent means (matched pairs) |
| Between Subjects ANOVA (one variable) | F tests | ANOVA: Fixed effect, omnibus, one-way |
| Between Subjects ANOVA (multiple variables) | F tests | ANOVA: Fixed effects, special, main effects and interactions |
| Repeated Measures ANOVA (multiple variables) | F tests | ANOVA: Repeated Measures, within factors |
| Mixed ANOVA (main effect) | F tests | ANOVA: Repeated Measures, between factors |
| Mixed ANOVA (interaction) | F tests | ANOVA: Repeated measures, within-between interaction |
| Multiple Regression (overall model) | F tests | Linear multiple regression: Fixed model, R2 deviation from zero |
| Multiple Regression (individual predictor) | F tests | Linear multiple regression: Fixed model, R2 increase |

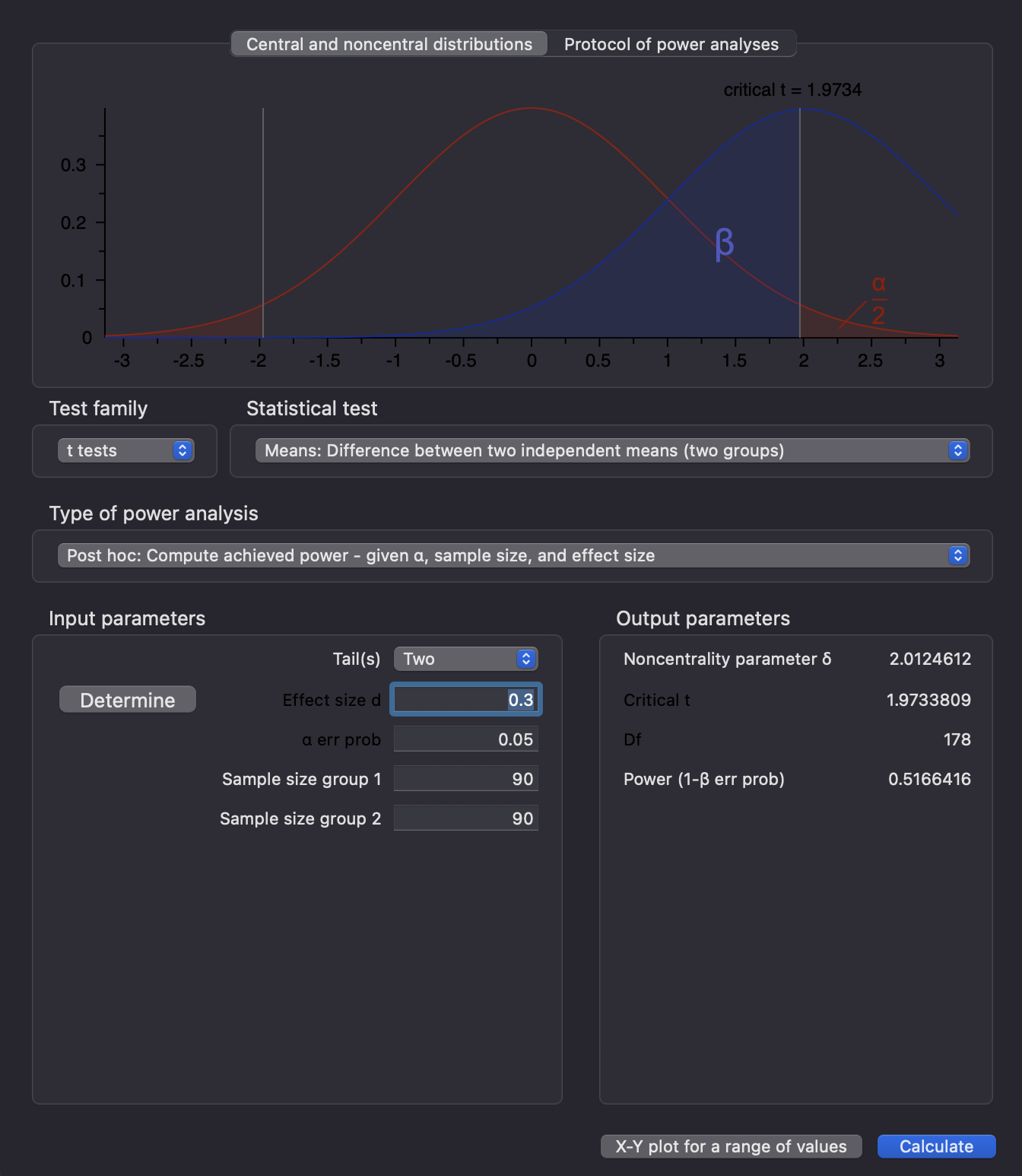
To demonstrate how these tools work, let’s calculate the power for our statistics classroom example. If we set our α to .05, power to .80, and our expected effect size to *d* = 0.30—our best guess for the predicted effect size based to previous research—, how many students would we need to recruit for the study? This study design is between-subjects, as each classroom consists of different individuals, and we have two groups. Therefore, we would select an independent *t*-test for our analysis, using a two-tailed test, as we did not make a prediction if the *R* or *JASP* class would do better. As shown in the output in Figure 9, we would need to enroll at least 352 students across the two groups in our study! The allocation ratio section is useful if you know one group will be larger than another, you can account for it here, otherwise leave as one. For example, if you know that the classes recruited for the *JASP* students will be larger than the *R* classes, you would take the ratio of general enrollment sizes (i.e., 40/25 = 1.6) and use that number in that box. One suggestion would be to estimate various possible effect sizes, power, and test type combinations to get a full picture of what may happen in your study.



*Figure 9*. A screen capture of G\*Power for estimating the required sample size for the first study on statistics classrooms. The bottom left panel would be filled in with the effect size, , power, and allocation ratio. The bottom right section indicates the number of participants required in each group and overall for these settings.

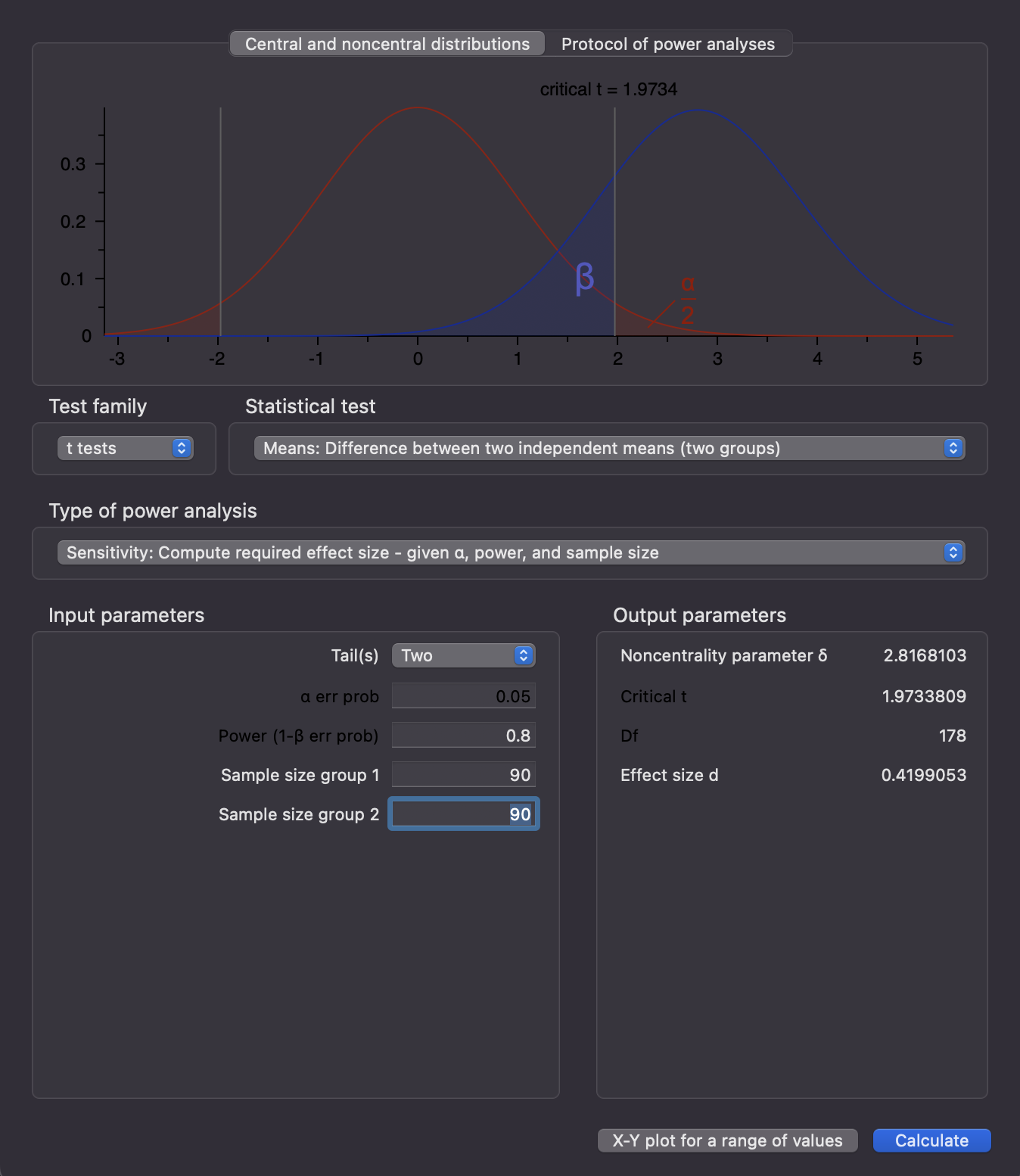
## Using Programs (Sensitivity)

What if we do not have the capability of recruiting that many students? The beauty of these programs is that we can switch from estimating the required sample size to determining the *potential* power or sensitivity of our study, given any research ability constraints. If you only have six classes with thirty students in class (i.e., 180 students total) per year for your experiment, what would your power be for the study with α set to .05 and an effect size set to *d* = 0.30? As shown in Figure 10, we would expect power to be approximately 52%—not very encouraging.



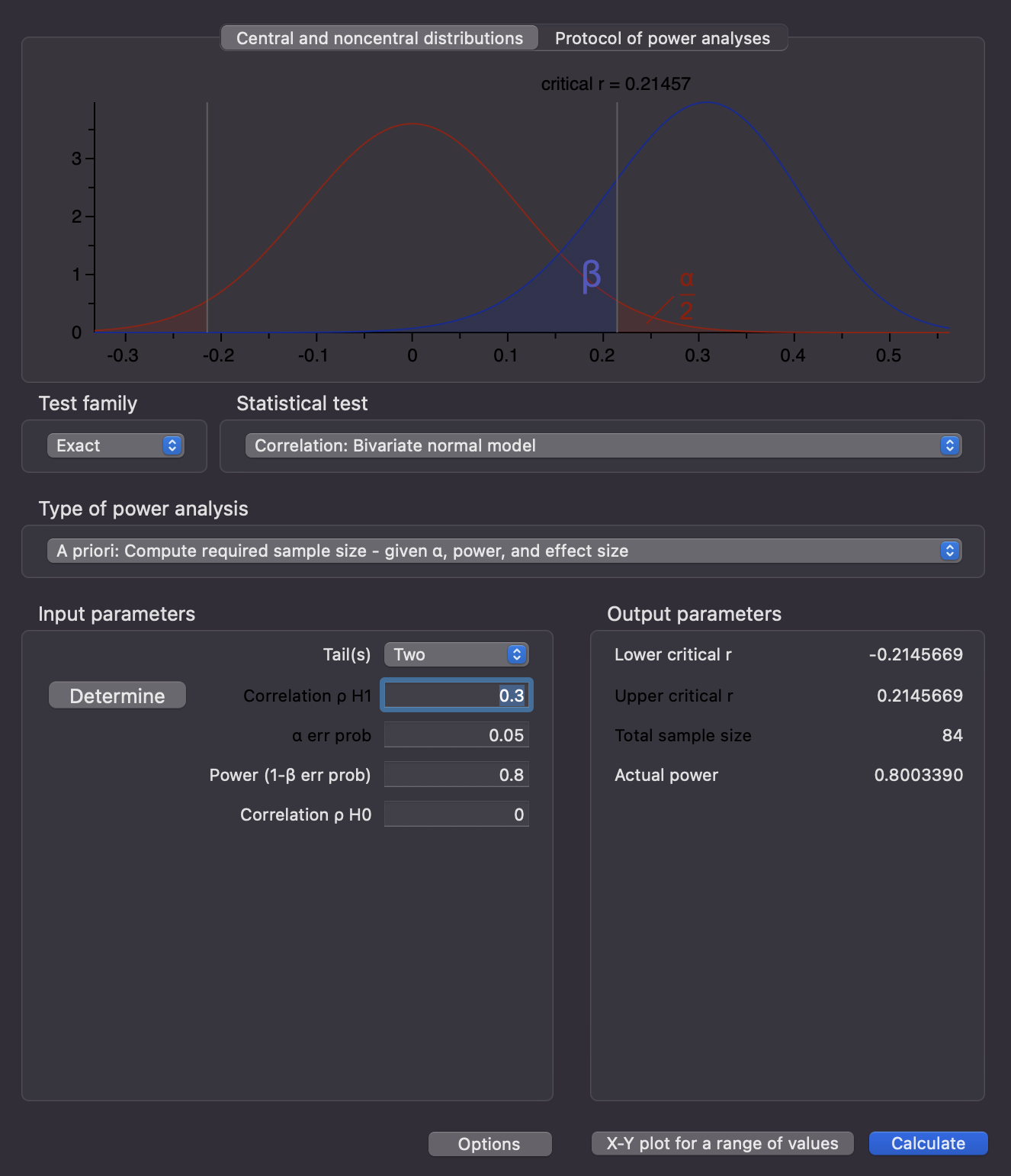
*Figure 10.* An estimation of prospective power (i.e., calculated before the study is conducted) for our first study on statistics classrooms using the same criterion as the previous power estimation but a specific sample size we know we can collect in the left-hand panel. The output is the likely power that can be detected with this sample size and effect size.

Instead, we may decide to calculate a sensitivity analysis that would indicate what effect size you could reasonably expect to detect with α set to .05 and power set to .80. When we recalculate, we find that the study would be able to detect an effect size of *d* = 0.42; this result is not extremely different from our estimated effect size (Figure 11). Given that we sample exactly 90 students in each group, the final study will not have the power to detect smaller effects.



*Figure 11.* An estimation of sensitivity for our first study on statistics classrooms using the same criterion as the previous power estimation but a specific sample size we know we can collect in the left hand panel. The output is the effect size that can be detected with that sample size and power values entered.

Our second study is also relatively straightforward to calculate in G\*Power using a correlation between data literacy and first job income of *r* = .30—often considered a medium effect size in social and behavioral sciences (Gignac & Szodorai, 2016). With this effect size and the same α and power we’ve used previously, we would need *n* = 84 participants to detect the effect shown in Figure 12. If we wanted to correlate other skills found on the graduation exit exam with first job income, we should determine the expected lowest effect size that may occur to ensure enough power to detect all the possible correlations individually.



*Figure 12.* Sample size estimation for study two examining the correlation between data literacy and income. In the bottom left panel, you would enter the correlation, , power, and the population correlation as comparison. The total sample size and power would appear in the bottom right-hand panel.

## Using *R* - Packages and Simulations

Finally, we turn to our last study that examines student data literacy and the scores on the game modules presented in class. These students were measured once a week for the entire semester—14 assessments excluding exam weeks. The study design is repeated measures, as we are measuring the same participants multiple times throughout the semester. As noted earlier, repeated measures designs usually have more power than the counterpart between subjects designs that makes them desirable to many researchers. However, the caveat to research designs with multiple measurement times is the statistical analyses can potentially increase in complexity along with the data collection. For example, no simple point and click program can adequately capture the intricacy of an ecological momentary assessment design; this complexity then leads to the necessity of statistical programming skills for both power planning and analysis of the data.

Before we estimate our sample size for the third study, let’s examine one of the most popular power planning packages available in *R*—*pwr* (Champely et al., 2017). The basic installation of *R* (called base *R*) comes with several functions and options for computing statistics and creating diagrams. The open-source community of programmers and researchers provide add-ons to base *R* (called packages) that allow you to customize your analyses. Nearly 20,000 *R* packages are available for download from the official R collection, and even more are accessible through GitHub—a collaborative, code-hosting platform. Using *pwr*, we can perform the same analyses as we did in G\*Power and reach the same conclusions about the sample size given the effect size, power, and α shown in Figure 13 and 14.

library(pwr)  
pwr.t.test(  
 n = NULL, # estimate N  
 d = 0.3, # effect size  
 sig.level = .05, # alpha  
 power = .80, # power level  
 type = "two.sample", # independent t is two.sample  
 alternative = "two.sided" # two tailed test  
)

##   
## Two-sample t test power calculation   
##   
## n = 175.3847  
## d = 0.3  
## sig.level = 0.05  
## power = 0.8  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

*Figure 13.* The code used from the *pwr* library in *R* to estimate the sample size for the classroom comparison study. Next to each code section, the purpose of the section is noted. The output notes that estimation is for each group separately, and therefore, we would have to multiply by the number of groups to get our final sample size.

pwr.r.test(  
 n = NULL, # estimate N  
 r = 0.3, # effect size  
 sig.level = .05, # alpha  
 power = .80, # power level  
 alternative = "two.sided" # two tailed test  
)

##   
## approximate correlation power calculation (arctangh transformation)   
##   
## n = 84.07364  
## r = 0.3  
## sig.level = 0.05  
## power = 0.8  
## alternative = two.sided

*Figure 14.* The code used from the *pwr* library to estimate the sample size needed for the correlation study on data literacy and job income.

Once our design becomes too complex for G\*Power, it often becomes too complex for most available programs, website applications, and *pwr*. In this scenario, researchers can simulate the study using other available packages in *R*. Simulation effectively involves *making up* plausible data that matches your study, estimating power at various sample and/or effect sizes, and then repeating this task multiple times to determine the average power at each combination of variables. The end result is sometimes a table much like the one we started this section with—the number of participants proposed, and the power found for each sample size. Thankfully, there are multiple packages that can be used to create fake data [e.g., *faux* DeBruine (2021) and *simstudy* Goldfeld and Wujciak-Jens (2020); see Figure 15].

set.seed(5839052)   
library(faux)  
library(nlme)  
# define parameters  
subj\_n = 10 # number of subjects  
item\_n = 7 # number of days, change from 7, 10, 14  
b0 = 0 # intercept  
b1 = 2 # fixed score for coefficients  
u0s\_sd = .5 # random intercept SD for subjects  
u0i\_sd = .5 # random intercept SD for days  
u1i\_sd = .5 # random b1 slope SD for days  
r01i = 0 # correlation between random effects 0 and 1 for days  
sigma\_sd = 2 # error SD  
  
sample\_sizes <- c(30, 40, 50, 60, 70, 80, 90, 100)  
p\_values <- data.frame(sample\_sizes = 1:(length(sample\_sizes)\*100),   
 p = 1:(length(sample\_sizes)\*100))  
  
row <- 1  
for (i in sample\_sizes){  
   
 for (r in 1:100){  
   
 # set up data structure  
 DF <- add\_random(subj = i, item = item\_n) %>%  
 # add both variables  
 add\_within("subj", measure = c("game\_1", "game\_2")) %>%   
 # add random effects   
 add\_ranef("subj", u0s = u0s\_sd) %>%  
 add\_ranef("item", u0i = u0i\_sd, u1i = u1i\_sd, .cors = r01i) %>%  
 add\_ranef(sigma = sigma\_sd) %>%  
 # calculate DV  
 mutate(dv = b0 + u0s + u0i + (b1 + u1i) + sigma) %>%   
 pivot\_wider(id\_cols = c("subj", "item"),   
 names\_from = "measure",   
 values\_from = "dv") %>%   
 mutate(positive = scale(positive))  
   
 # save study results   
 save <- summary(lme(negative ~ positive,   
 data = DF,   
 random = list(~1|subj)))  
 p\_values[row, ] <- c(i, save$tTable[2, 5])  
   
 row <- row + 1  
   
 }  
   
}  
  
# calculate power   
day\_data <- p\_values %>%  
 group\_by(sample\_sizes) %>%   
 summarize(Power = sum(p <=.05)/100)  
  
# export data to save   
export(day\_data, "seven\_days.csv", row.names = F)

*Figure 15.* The code for simulating the data and estimating power for our repeated measures study on data literacy and gamification.

*Table 3.* Simulation estimation of power for several potential sample sizes and number of measurement days for the gamification of data literacy study.

| Sample Size | 7 Days | 10 Days | 14 Days |
| --- | --- | --- | --- |
| 30 | 0.55 | 0.59 | 0.72 |
| 40 | 0.52 | 0.65 | 0.79 |
| 50 | 0.68 | 0.78 | 0.89 |
| 60 | 0.61 | 0.87 | 0.92 |
| 70 | 0.75 | 0.88 | 0.96 |
| 80 | 0.69 | 0.88 | 0.94 |
| 90 | 0.86 | 0.92 | 1.00 |
| 100 | 0.83 | 0.97 | 0.97 |

Our study on gamification in the classroom not only includes multiple assessments measuring participants across the semester but also includes multiple variables using multiple game module scores to predict data literacy. As shown in Table 3, we could simulate the power of having multiple game modules predicting data literacy at a medium effect size for sample sizes of 30 to 100 participants. We also simulated the number of weeks we might need to measure participants to achieve a certain level of power (7, 10, and 14 weeks). As noted in our type of test section, repeated measures designs often show more power than between subjects designs, and each subsequent measurement may also increase power. For each sample size, the results indicate that we increase our power by increasing the number of days we measure each participant. We can use a simulation test similar to this one to manipulate multiple factors that influence our study (sample size, measurement days) at once and compare what might be the best scenario for our project. Participants may drop out of the study if it becomes too long without incentive, so we could compromise to collect *n* = 60 participants over ten days to achieve at least 80% power. Table 3 shows the estimated power and this value appears to represent the best compromise of number of days and participants needed.

With this example, hopefully you can see how complex power can be to estimate. The goal is often to pinpoint an appropriate sample size for a research study, or the sensitivity of the study given a fixed sample size. The desire would be to have one set estimate when planning the study, but as one considers all the potential influences on power, the number of “right” answers for power estimation can increase substantially. Estimations are often our best guess by fixing a known set of variables (α, effect size, study type, etc.) that may not represent the true values found in the population of study. Estimating across multiple best guesses can provide you with a range of potential outcomes to consider.

## Newer Methods - AIPE

In light of these considerations, a newer method of power and sample size planning involves accuracy in parameter estimation [AIPE; Maxwell et al. (2008); Kelley (2007); Kelley et al. (2018)]. Power, traditionally defined, focuses on finding a significant effect—that is a *p*-value less than your set cut-off criterion (i.e., α). Not all studies involve a specific hypothesis test to use in the power analysis; often, studies include multiple hypothesis tests, complicated statistics that may not fit into point and click programs, or simply no hypothesis test at all [e.g., a collection for a database of values that other researchers will use for their study planning or exploratory analyses; Buchanan et al. (2019)]. Instead, using AIPE, researchers can estimate sample size or sensitivity of their planned design to create an “accurately measured” parameter of interest. Parameters are simply something you want to measure (e.g., means, effect size, or regression slopes).

But what does “accurately measured” actually mean? First, a researcher would specify the parameter they want to measure and estimate from their study. Next, they would decide what a “sufficiently narrow” window of measurement around that parameter would be by using confidence intervals. Confidence intervals are a range of values around a parameter that represent an estimate of the true population parameter; they help us understand the uncertainty in our measurement. For example, the mean in our study one *R* class could be 82 points, while the mean for the *JASP* class is 78 points. It would appear that the *R* class did better! However, when we add a confidence interval around the mean for each class, we can see that each has a good bit of variability, and they likely are not that different—*R*: 95% CI [75, 89] and *JASP*: 95% CI [68, 88]. In fact, these two confidence intervals show that the scores in the classrooms are very similar because they overlap each other. The percentage (95%) presented with the confidence interval represents the percent of time that one might expect the true population mean to be between the lower and upper numbers.

Therefore, the researcher would define the amount of variability (i.e., the width of the confidence interval around their parameter they are willing to accept) as “sufficiently narrow”. Conveniently, the *MBESS* package in *R* can make these calculations straightforward for researchers (Kelley, 2022) with the caveat that it can be tricky to know exactly what width to use for confidence intervals, as this procedure is fairly new to most researchers. This area can be paired with determining the smallest effect size of interest [i.e., the smallest effect size a researcher would determine as practically important or useful for their study; Anvari and Lakens (2021); Riesthuis et al. (2022)]. Previously, we used an effect size of *d* = 0.30 for our study examining the differences in statistics classrooms . We may decide that the smallest effect of interest is *d* = 0.10—a small change in grades when using one statistical program over another. If the study shows anything smaller than that effect size, we would suggest the difference between classrooms was not useful enough to make instructors switch to a new instruction method. Using *MBESS*, we would find that the suggested sample size is over 700 participants (Figure 16)!

library(MBESS)  
ss.aipe.smd(delta = 0.30, # the effect size  
 conf.level = 0.95, # confidence interval   
 width = 0.20) # our suggested width

## [1] 777

*Figure 16*. AIPE estimation code and sample size output from *MBESS* used to estimate the number of participants needed to find a *d* value between 0.10 and 0.50 (0.30 0.20 width).

We can similarly use this procedure to estimate how many participants would be necessary for our second study on data literacy and job income. We previously used *r* = .30 and could similarly suggest that we would not be interested if the correlation was below *r* = .10. This sample size is also very large in comparison to our previous power estimations—over 2,800 participants (see Figure 17).

ss.aipe.R2(Population.R2 = 0.30^2, # our correlation squared  
 conf.level = 0.95, # confidence interval   
 width = 0.20^2, # the width squared  
 p = 1) # one predictor X to Y

## [1] "The approximate sample size is given below; you should consider using the additional"  
## [1] "argument 'verify.ss=TRUE' to ensure the exact sample size value is obtained."

## $Required.Sample.Size  
## [1] 2863

*Figure 17.* AIPE estimation code and sample size output for the correlation study. In this example, we use —the estimated correlation r squared—, and we similarly square the width of the confidence interval we are looking for.

One thing to remember is that we used a power level of 80% in the previous estimations, yet the AIPE procedure generally uses a confidence interval of 95%. If we recalculate the power at 95% for our first study, we find a much larger number of 580 participants total (290 for each group) in Figure 18 (as compared to 777 from *MBESS* in Figure 16 and 350 total from *pwr* at 80% power in Figure 13). However, the sample size estimate for our correlation study is much smaller in sample size estimate for 95% power at 138 participants in Figure 19 (compared to *MBESS* estimation of 2863 in Figure 17 and 84 from *pwr* at 80% power in Figure 14). These two estimation methods are not meant to give the same answer, as they have different goals to achieve (i.e., a specific *p*-value versus a specific confidence interval). One caution with using accuracy in parameter estimation, though, is that while AIPE is advertised as a way to estimate sample size regardless of hypothesis test, confidence intervals still use the framework of null hypothesis significance testing. Therefore, a 95% confidence interval for a correlation *r* of [.10, .50] would also usually be “significant” using *p*-value criterion and α < .05 because the confidence interval does not include zero (i.e., the estimation of the true population *r* does not include zero because the entire confidence interval is greater than zero).

pwr.t.test(  
 n = NULL, # estimate N  
 d = 0.3, # effect size  
 sig.level = .05, # alpha  
 power = .95, # power level  
 type = "two.sample", # independent t,   
 alternative = "two.sided" # two tailed test  
)

##   
## Two-sample t test power calculation   
##   
## n = 289.7353  
## d = 0.3  
## sig.level = 0.05  
## power = 0.95  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

*Figure 18.* The code for 95% power using the *pwr* package in *R* for our statistics classroom example. In comparison to our previous 80% power, we find that we need more people; this matches the AIPE estimations.

pwr.r.test(  
 n = NULL, # estimate N  
 r = 0.3, # effect size  
 sig.level = .05, # alpha  
 power = .95, # power level  
 alternative = "two.sided" # two tailed test  
)

##   
## approximate correlation power calculation (arctangh transformation)   
##   
## n = 137.7587  
## r = 0.3  
## sig.level = 0.05  
## power = 0.95  
## alternative = two.sided

*Figure 19.* The code for 95% power for the data literacy and job income correlation example. The AIPE estimations are still much larger than the estimations using normal power approximations.

# Limitations

## What Power Does Not Say

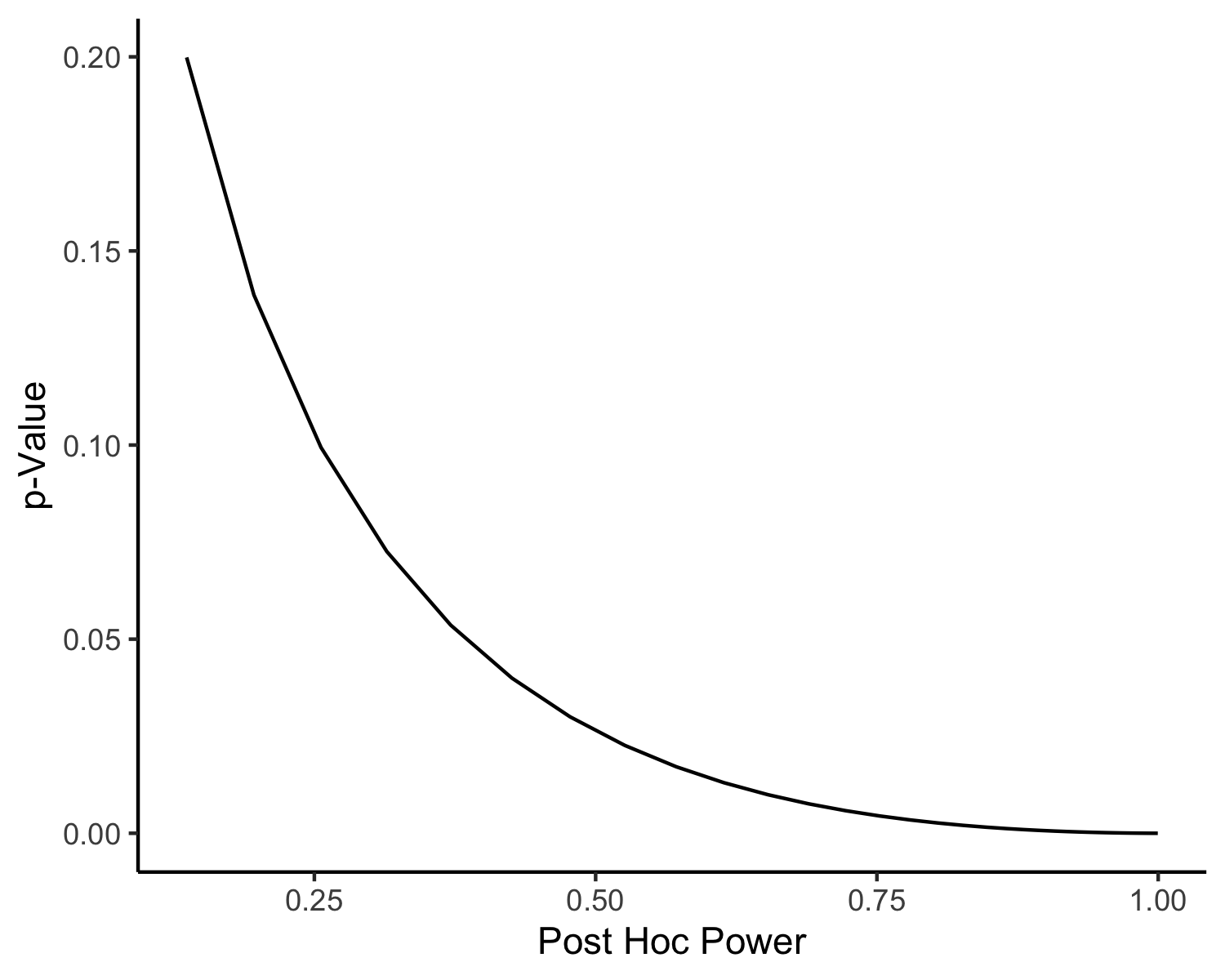
Let’s go back to the definition of power. It is the likelihood of detecting a significant effect, given the effect size, α, sample size, and so on, ***if that effect exists*** (e.g., if the null hypothesis is actually false). When using NHST, we usually want the null hypothesis to be false, (i.e., supporting our research hypothesis) . Power does not tell you if the null hypothesis is in reality false—just the likelihood of rejecting the null if it is. We often do not know if the null is truly false (and therefore, we should move on from our study idea); this ambiguity can create a scenario that you might think the effect exists, but maybe you did not have enough power to detect it (and continue to look for something that does not exist). This scenario happens to researchers quite a bit—samples cost time, money, and effort to collect. You cannot always get as many data points as you would like, and even if you do, the other factors estimated within a power analysis may be the wrong guess.

If you overestimate your effect size, you likely will not have enough power to find the smaller effect size that truly exists in the population. Additionally, it is important to remember publication bias—the published background research likely overestimates the true effect size as only the “significant” findings will be published (Franco et al., 2014). Last, effect sizes *also* have variance, just like any other statistical parameter (e.g., the mean has a standard deviation). Therefore, when a researcher runs a study, even if they estimated the right effect size for their power analysis, they could find something higher or lower depending on their particular sample (Pek et al., 2022). Thus, even with our best intentions, power analyses do not guarantee that we will detect a true effect and estimates that we used in the power analysis often differ from the final study values.

## Post Hoc Power

As you might have noticed, this chapter has mainly focused on calculating the necessary sample size to plan a study or the sensitivity of a study with a specific sample size. In practice, we could calculate the amount of power found in a study after it is complete [i.e., *post hoc* power or retrospective power; Dziak et al. (2020)]. Reviewers of research papers may request this information to know if the study was “powered appropriately” for a non-significant result. However, once the data is collected, we know if we have found a significant effect or not, and any *post hoc* power analysis will reflect the same answer as the *p*-value from the study (i.e., if the result is non-significant, the power analysis will say there was not enough power).

For example, let’s examine our study focusing on job income and data literacy using a correlation of *r* = .30. If we only had twenty participants, we might find that *post hoc* power was .26. If we collected 100 participants, then power jumps to .86. This analysis seems like it says something useful right? However, *post hoc* power is only a representation of the *p*-value found in the study, which here, was influenced by the change in sample size. Therefore, all we have learned is something we knew from earlier in this chapter: sample size has a large influence on power. Figure 20 shows this relationship for our *r* = .30. As sample size increases, power increases, and *p*-values decrease. Therefore, if you had 300 participants, the *post hoc* power would be high and the *p*-value would be low, which does not tell you much about the actual ability to detect your effect in a new study. Instead, one should provide a sensitivity analysis or simply the effect size found in the study with the 95% confidence interval to understand the potential range of the effect found with the study parameters (Heckman et al., 2022).



*Figure 20*. The relationship between *post hoc* power and *p*-values for *r* = .30. As shown, the power increases exponentially with decreasing *p*-values.

# Suggestions

Given all that you’ve learned in this chapter, what guidelines should you use when performing power analyses?

* Use previous literature in your study area to estimate possible effect sizes; however, keep in mind that not all studies are published, and the potential effect size may be only half the size of what is published (Open Science Collaboration, 2015). Use the smallest effect size you find to help overestimate sample size for adequate power.
* Select and justify the choices of , , and power when planning your research study.
* If possible, calculate necessary sample size from several sources (calculators, programs, simulation, AIPE) and with several parameters (different effect sizes, study manipulations) to explore the potential power outcomes for your study.
* Try to achieve the largest sample size suggested by different power analyses, as sample size is one of the largest influences on power.
* Consider multiple types of research designs and statistical analyses for their impact on the power of the study.
* When sample size is predetermined by resources, calculate the sensitivity of your study to detect various effect sizes.

# References

Aerts, M., Molenberghs, G., & Thas, O. (2021). Graduate education in statistics and data science: The why, when, where, who, and what. *Annual Review of Statistics and Its Application*, *8*(1), 25–39. <https://doi.org/10.1146/annurev-statistics-040620-032820>

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. <https://doi.org/10.1177/0956797617723724>

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. <https://doi.org/10.1016/j.jesp.2021.104159>

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. <https://doi.org/10.1038/s41562-017-0189-z>

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). LAB: Linguistic annotated bibliography a searchable portal for normed database information. *Behavior Research Methods*, *51*(4). <https://doi.org/10.3758/s13428-018-1130-8>

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H. (2017). *Pwr: Basic functions for power analysis*.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. <https://doi.org/10.4324/9780203771587>

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. <https://doi.org/10.1038/d41586-022-00150-2>

Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., & Edlund, J. E. (2021). What Crowdsourcing Can Offer to Cross-Cultural Psychological Science. *Cross-Cultural Research*, *55*(1), 3–28. <https://doi.org/10.1177/1069397120950628>

*Data Scientists : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. (n.d.). <https://www.bls.gov/ooh/math/data-scientists.htm>

DeBruine, L. (2021). *Faux: Simulation for factorial designs*. Zenodo. <https://doi.org/10.5281/ZENODO.2669586>

Dienes, Z. (n.d.). *Testing theories with bayes factors* (A. L. Nichols & J. E. Edlund, Eds.; 1st ed., Vol. 1).

Dienes, Z. (2008). *Understanding psychology as a science*. Red Globe Press London. <https://link.springer.com/book/9780230542303>

Dziak, J. J., Dierker, L. C., & Abar, B. (2020). The interpretation of statistical power after the data have been gathered. *Current Psychology*, *39*(3), 870–877. <https://doi.org/10.1007/s12144-018-0018-1>

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. <https://doi.org/10.3758/BF03203630>

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

Goldfeld, K., & Wujciak-Jens, J. (2020). Simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, *5*(54), 2763. <https://doi.org/10.21105/joss.02763>

Gümüş, M. M., & Kukul, V. (2023). Developing a digital competence scale for teachers: validity and reliability study. *Education and Information Technologies*, *28*(3), 2747–2765. <https://doi.org/10.1007/s10639-022-11213-2>

Heckman, M. G., Davis, J. M., & Crowson, C. S. (2022). Post hoc power calculations: An inappropriate method for interpreting the findings of a research study. *The Journal of Rheumatology*, *49*(8), 867–870. <https://doi.org/10.3899/jrheum.211115>

JASP Team. (2022). *JASP (version 0.16.3)[computer software]*. <https://jasp-stats.org/>

Kelley, K. (2022). *MBESS: The MBESS r package*. <https://CRAN.R-project.org/package=MBESS>

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*(4), 755–766. <https://doi.org/10.3758/BF03192966>

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. <https://doi.org/10.1037/met0000127>

Koch, C., & Jones, A. (2016). Big science, team science, and open science for neuroscience. *Neuron*, *92*(3), 612–616. <https://doi.org/10.1016/j.neuron.2016.10.019>

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. <https://doi.org/10.3389/fpsyg.2013.00863>

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Calster, B. V., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. <https://doi.org/10.1038/s41562-018-0311-x>

Legaki, N.-Z., Xi, N., Hamari, J., Karpouzis, K., & Assimakopoulos, V. (2020). The effect of challenge-based gamification on learning: An experiment in the context of statistics education. *International Journal of Human-Computer Studies*, *144*, 102496. <https://doi.org/10.1016/j.ijhcs.2020.102496>

Matli, W., & Ngoepe, M. (2020). Capitalizing on digital literacy skills for capacity development of people who are not in education, employment or training in South Africa. *African Journal of Science, Technology, Innovation and Development*, *12*(2), 129–139. <https://doi.org/10.1080/20421338.2019.1624008>

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115. <https://doi.org/10.1086/288135>

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*(1), e0208631. <https://doi.org/10.1371/journal.pone.0208631>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. <https://doi.org/10.1126/science.aac4716>

Pangrazio, L., Godhe, A.-L., & Ledesma, A. G. L. (2020). What is digital literacy? A comparative review of publications across three language contexts. *E-Learning and Digital Media*, *17*(6), 442–459. <https://doi.org/10.1177/2042753020946291>

Pek, J., Pitt, M. A., & Wegener, D. T. (2022). Uncertainty limits the use of power analysis. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001273>

Porras-Segovia, A., Molina-Madueño, R. M., Berrouiguet, S., López-Castroman, J., Barrigón, M. L., Pérez-Rodríguez, M. S., Marco, J. H., Díaz-Oliván, I., de León, S., Courtet, P., Artés-Rodríguez, A., & Baca-García, E. (2020). Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *Journal of Affective Disorders*, *274*, 733–741. <https://doi.org/10.1016/j.jad.2020.05.067>

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Ralston, K. (2020). ‘Sociologists Shouldn’t Have to Study Statistics’: Epistemology and Anxiety of Statistics in Sociology Students. *Sociological Research Online*, *25*(2), 219–235. <https://doi.org/10.1177/1360780419888927>

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, *36*(1), 203–215. <https://doi.org/10.1002/acp.3911>

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31. <https://doi.org/10.1016/j.ijhcs.2014.09.006>

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., & others. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337356.

*Statistical rethinking: A bayesian course with examples in R and STAN*. (n.d.). <https://www.routledge.com/Statistical-Rethinking-A-Bayesian-Course-with-Examples-in-R-and-STAN/McElreath/p/book/9780367139919>

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. <https://doi.org/10.1177/1745691617751884>

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*, 428–443. <https://doi.org/10.1037/1082-989X.10.4.428>