Spring\_SemesterProj

Soyeon Lee

5/20/2021

## Overview

Power analysis performed during the planning stage of the research study is a valuable tool; it prevents wasting time and resources invested in the study. Power is the probability that a study will reject the null hypothesis when it is in fact, false. In the context of the social and behavioral science, power is especially useful in finding the treatment or intervention effect. High power means the study design can detect the effect with a higher probability; whereas lower power means the study design will miss the important but small effect.

In this write-up, we will identify the function of power in a research study, its determinants, and demonstrate the application of power.

## Concept

Importance power in hypothesis testing There are two purposes of statistics in social and behavioral sciences: describe the observed phenomenon and explain how the results of samples reveal a phenomenon in a broader context. When a research study involves the latter, we use samples to conduct experiments to learn about the population. Samples are meant represent the population of interest, but they may not represent the population well. In other words, results of the samples differing from the estimated values of population may be due the random fluctuation of the samples, rather than indicating the real effects of the treatment or intervention.

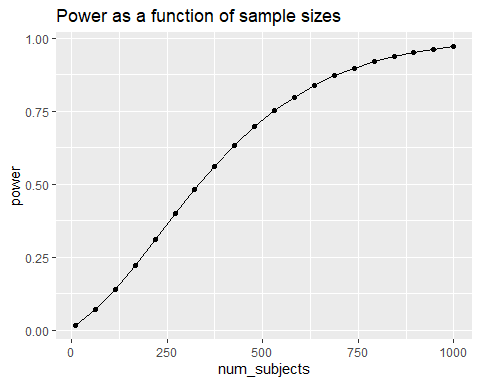
When we conduct hypothesis test to compare the null against the alternative hypothesis, power can be useful. There are two types of errors that can occur when we decide to accept or reject the null hypothesis: Type I error and type II error. We commit type I error when we erroneously reject the null hypothesis that the population data without treatment or intervention do not differ from the population with treatment or intervention. Type II error is when we fail to reject the null hypothesis when the alternative hypothesis is true. That is, we decide that any deviation of the values after intervention or treatment is due to the treatment effect, rather than random fluctuation of the samples, or sampling error. Power analysis concerned with the Type II error; it is the probability that we will correctly detect the effect if there is one.

Although useful to know for designing a research study, we do not know the level of power in advance. The probability of correctly rejecting the null hypothesis vary with sample size, effect size, and criterion for rejection. We also do not know the three determinants in advance; however, few considerations should be made.

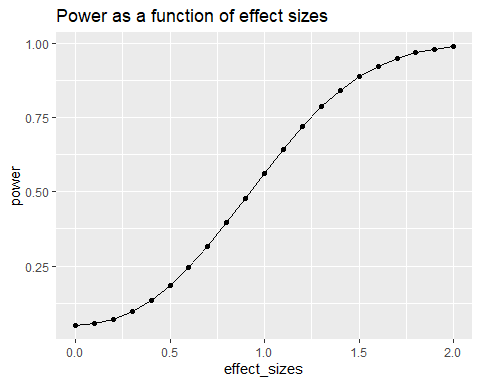
### Sample size

Power varies with size of samples. Given that estimated standard deviation is the same, larger sample size decreases sampling error. Reduced sampling size is optimal, as it increases precision of statistical tests to distinguish treatment effects from differences occurring due to sampling error. Small sample sizes leading to unstable results that are more prone to Type II error (i.e., missing the treatment effect). With other things constant, power varies with sample size.

library(pwr)  
num\_subjects <- seq(10,1000,length.out=20)  
power <- sapply(num\_subjects,  
 FUN = function(x) {  
 pwr.t.test(n=x,  
 d=.2,  
 sig.level=.01,  
 type="two.sample",  
 alternative="two.sided")$power})  
plot\_df <- data.frame(num\_subjects,power)  
  
library(ggplot2)  
ggplot(plot\_df, aes(x=num\_subjects,  
 y=power))+  
 geom\_point()+  
 geom\_line()+  
 ggtitle("Power as a function of sample sizes")

 ### Effect size One of the most common effect size measured is Cohen’s d, which is a standardized mean difference indicating how much impact treatment has on the dependent variable. Effect sizes are classified as small, medium, and large–approximately .2, .5, and .8, respectively. All other things equal, large effect size makes it easy to detect treatment effect; whereas a very small effect size may not reliably detect the treatment effect. In other words, power increases as effect size increases.

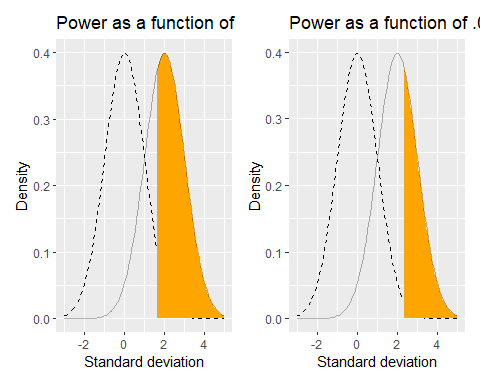
library(pwr)  
effect\_sizes <- seq(0,2,.1)  
power <- sapply(effect\_sizes,  
 FUN = function(x) {  
 pwr.t.test(n=10,  
 d=x,  
 sig.level=.05,  
 type="two.sample",  
 alternative="two.sided")$power})  
plot\_df <- data.frame(effect\_sizes,power)  
  
library(ggplot2)  
ggplot(plot\_df, aes(x=effect\_sizes,  
 y=power))+  
 geom\_point()+  
 geom\_line()+  
 ggtitle("Power as a function of effect sizes")

 ### Decision criterion When setting the alpha criterion, the goal is to minimize Type I errors. In other words, a researcher often determines the criterion value with the interest of not rejecting the null when they are actually true. Alpha criterion of .05 rather than .01 more often used. This lenient standard (i.e., larger alpha criterion) is easier to reject the null hypothesis. With other things equal, larger alpha criterion increases the probability of rejecting the null hypothesis, when the alternative hypothesis is true. As demonstrated in the plot below, power (orange region) is greater when the alpha criterion value is larger.

library(ggplot2)  
library(patchwork)

## Warning: package 'patchwork' was built under R version 4.0.4

#graph for alpha=.05, one-tailed  
A<-ggplot(data.frame(x = c(-3, 5)), aes(x)) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "line",  
 linetype = 2  
 ) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "area",  
 fill = "orange",  
 args = list(  
 mean = 2  
 ),  
 xlim = c(1.64, 5)  
 ) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "line",  
 linetype = 1,  
 alpha = .3,  
 args = list(  
 mean = 2  
 )  
 ) +  
   
 labs(  
 title = "Power as a function of .05 criterion",  
 x = "Standard deviation",  
 y = "Density"  
 )  
  
#graph for alpha=.01, one-tailed  
B<-ggplot(data.frame(x = c(-3, 5)), aes(x)) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "line",  
 linetype = 2  
 ) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "area",  
 fill = "orange",  
 args = list(  
 mean = 2  
 ),  
 xlim = c(2.33, 5)  
 ) +  
   
 stat\_function(  
 fun = dnorm,  
 geom = "line",  
 linetype = 1,  
 alpha = .3,  
 args = list(  
 mean = 2  
 )  
 ) +  
   
 labs(  
 title = "Power as a function of .01 criterion",  
 x = "Standard deviation",  
 y = "Density"  
 )  
  
A+B



## Practical

Power analysis is useful both planning and diagnosis of research study design. During planning, it is frequently used to determine the sample size for the alpha criterion significance. As a diagnostic tool, researchers can identify if there is a enough design for power in the design, or learn about the reliable effect size to achieve a certain level of power. Such versatility is due to that power is a function of sample size, effect size, and alpha criterion.

In this section, we will walk through the example cases in which we will discover the appropriate sample and effect sizes to achieve power in an experiment. Note that the power of .80 is a conventionally used, which we adopt in our examples. This value can vary for specific purposes, but the use of the convention seems reasonable, according to documented evidence. The value of alpha criterion is more arbitrary but in our example, we will explore cases with .05 alpha criterion.

### Case 1: Determining the sample size for an experiment

As discussed in the concept section, large number of samples increases power, which can cause concerns about the Type I error (i.e., reject the null when there is no effect). On the other hand, too little power can cause researchers to make Type II error of failing to reject the null and miss the true effect. This trade-off between two types of errors warrants a power analysis to determine the sample size. We would like to find out the range of sample sizes that would help us reach the .80 level of power. Learning about the minimum sample size would be helpful to design an experiment without wasting additional time and resource.

Let’s say we are designing an experiment on the efficacy of a study skills training. The experimental group will receive an 8-week training and the controls will receive none. The test scores will then be compared between groups. Based on the previous literature, we estimate that the effect size of the training is d=.40. We want to achieve the power of .80, with the alpha criterion of .05. What would be the range and the minimum number of sample size required to detect that the training makes a difference in the students’ test performance?

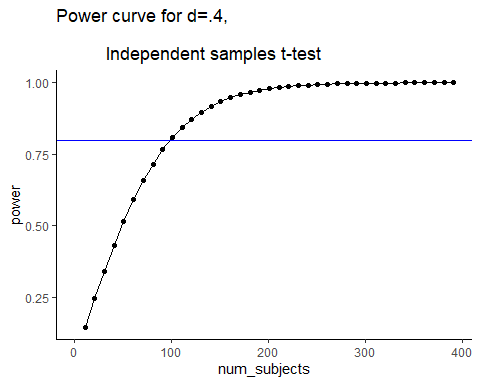
library(pwr)  
num\_subjects <- seq(1,400,10)  
power <- sapply(num\_subjects,  
 FUN = function(x) {  
 pwr.t.test(n=x,  
 d=.4,  
 sig.level=.05,  
 type="two.sample",  
 alternative="two.sided")$power})

## Warning in qt(sig.level/tside, nu, lower = FALSE): NaNs produced

plot\_df <- data.frame(num\_subjects,power)  
  
library(ggplot2)  
ggplot(plot\_df, aes(x=num\_subjects,  
 y=power))+  
 geom\_point()+  
 geom\_line()+  
 theme\_classic ()+  
 geom\_hline(yintercept=.8, color="blue")+  
 ggtitle("Power curve for d=.4, \n  
 Independent samples t-test")

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 row(s) containing missing values (geom\_path).

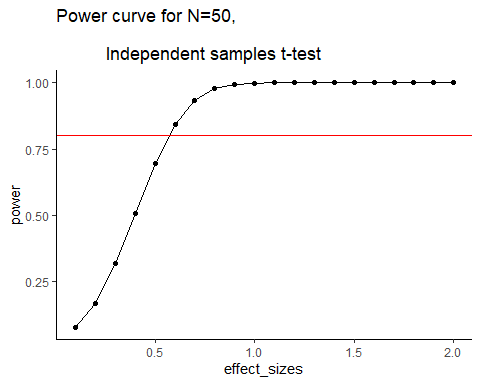
 According to the power analysis, the range of subject numbers appropriate for the goal of the experiment is around 100 participants and above per group. In consideration of resources and time limitation, designing the study with the minimum number of participants would be ideal.

### Case 2: Determining the effect size for an experiment

What if we have to find out the required effect size, instead of the number of participants? If we were to have limited number of participants already set for the experiment, what would be the range and the minimum effect size to result in statistically significance?

Let’s say we have 50 participants available per group. We want to define significance at .05 alpha criterion and achieve .80 level of power. What kind of effect size should we have in order to detect the intervention effect if there is one?

effect\_sizes <- seq(.1,2,.1)  
power <- sapply(effect\_sizes,  
 FUN = function(x) {  
 pwr.t.test(n=50,  
 d=x,  
 sig.level=.05,  
 type="two.sample",  
 alternative="two.sided")$power})  
plot\_df <- data.frame(effect\_sizes,power)  
  
library(ggplot2)  
ggplot(plot\_df, aes(x=effect\_sizes,  
 y=power))+  
 geom\_point()+  
 geom\_line()+  
 theme\_classic ()+  
 geom\_hline(yintercept=.8, color="red")+  
 ggtitle("Power curve for N=50, \n  
 Independent samples t-test")

 Looking at the result, we can see that the effect sizes of more than d=.5 (medium effect) would be need to achieve the experimental goal.

Power analysis is versatile because power is a function of the effect size, sample size, and the decision criterion for determining significance. If you know any one of four values (including power), you can solve for the remaining three values. As introduced in the example above, we can use the conventional values such as alpha criterion of .05 and power level of .80 in determining other values. That said, these conventions are not obligatory for researchers but useful. The level of power at .80 is a reasonable convention supported well by previous literature. For example, correctly rejecting the false null hypothesis is as four times as likely for the power of .80.

Though power analysis is seemingly useful in research designs, we must consider the related complexities. Achieving high power is not easy. For example, the level of power is often small in social and behavioral science. This may be related to many factors including the fields’ vulnerabilities to resources and luck to achieve a large sample size. Whereas larger effect sizes increases power, its flexibility is limited. In fact, they are always estimates, which are substantially based on theoretical and empirical research in many fields. Further limiting the ability to adjust the determining values for power is alpha criterion–specifically, the trade-off between Type I and Type II error imposes a limit to adjusting the decision criterion more liberal. In spite of these challenges, thoughtful application of power analysis in any statistical hypothesis testing is useful to justify the researchers’ rigorous effort.

## Generalization Problem

1a. Let’s say we want to re-design the previously mentioned study skills training. We think that there will be too much individual differences among available participants, such that we want to see whether students’ performance will be improve if they receive the 8-week training. We plan to have each participant to take a test before the training and after the training. Based on the previous literature, we estimate that the effect size of the training is d=.40. We want to achieve the power of .80, with the alpha criterion of .05. What would be the minimum number of sample size required to detect the improvement the training would make on the students’ test performance? Plot to show the range of values for sample sizes for this design.

1b. Compared to the required sample size for case 1, what are some of the reasons why significantly less number of participants could potentially achieve the same level of power?

### Suggested solutions

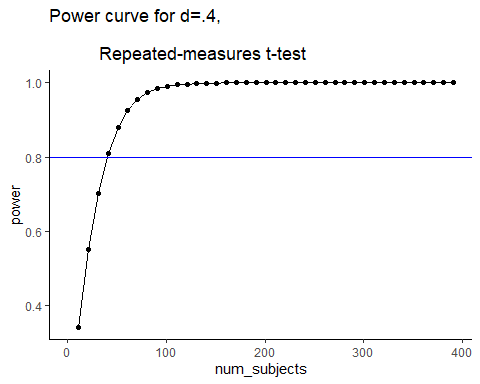
library(pwr)  
num\_subjects <- seq(1,400,10)  
power <- sapply(num\_subjects,  
 FUN = function(x) {  
 pwr.t.test(n=x,  
 d=.4,  
 sig.level=.05,  
 type="paired",  
 alternative="greater")$power})

## Warning in qt(sig.level/tside, nu, lower = FALSE): NaNs produced

plot\_df <- data.frame(num\_subjects,power)  
  
library(ggplot2)  
ggplot(plot\_df, aes(x=num\_subjects,  
 y=power))+  
 geom\_point()+  
 geom\_line()+  
 theme\_classic ()+  
 geom\_hline(yintercept=.8, color="blue")+  
 ggtitle("Power curve for d=.4, \n  
 Repeated-measures t-test")

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 row(s) containing missing values (geom\_path).

 1a solution: Almost 40 participants at minimum are required to detect that the detect the improvement the training would make on the students’ test performance.

1b solution: Significantly reduced number of participants are required to achieve same level of power in this experiment because reduced sampling error would increases the probability of detecting the intervention effect. Additionally, one-tailed test of this current example lowers the boundaries for critical region, which in turn increases power.

## References