

Public Policy 529

Power Analysis

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

December 4, 2023

Outline

1. Preliminaries
2. Digging In
3. Various Examples
4. Summary

Outline

1. Preliminaries

2. Digging In

3. Various Examples

4. Summary

What is Power Analysis?

- We seek to determine the minimum sample needed to detect an effect at a particular level of significance.
- This is important at the stage of research design. We are about to undertake data collection, how much data do we need?
- Power analysis can also determine how large an effect size needs to be in order for us to detect it with a given sample size.
- We have to think ahead to the hypothesis testing stage and anticipate what will be necessary to reject H_0 .

Statistical Power Defined

- The power of a test is the probability we will reject H_0 if H_A is correct.
- It is the complement of an error of Type II (β), which is the probability that we will fail to reject H_0 when it is false.
- Specifically, statistical power is $1 - \beta$.
- For minimum desired power, a conventional level is .8. If H_A is correct, there is a probability of .8 that we will reject H_0 .
- Higher levels may be preferable.

Possible Results from a Hypothesis Test

Truth	Decision	
	Reject H_0	Do not Reject H_0
H_0 True	Type I error (α)	Correct
H_0 False	Correct	Type II error (β)



Power Depends on the Following

- **Sample size**: larger samples decrease the standard error and make it more likely we would reject an incorrect H_0 .
- **Effect size**: the larger the effect size, the more likely we will reject H_0 because it is farther away.
- **α -level**: the smaller is α , the less likely we are to reject H_0 even if it is false.
- If we know any two of these things, we can calculate the third when given a particular level of power.

But Some of These Things are Unknown

- Obviously, we may not know the effect size in advance of doing the study.
- We can, however, **decide upon a minimum effect size** that is substantively meaningful and seek enough power to detect it.
- There are other things we can't know in advance, such as sample standard deviations, which are needed for standard errors.
- We may have to estimate these things based on previous data or do a pre-test to get some information.

Measuring Effect Size (E)

- In power analysis, the effect size is standardized, meaning that the original scale is divided by a standard deviation.
- So, for the difference between two means:

$$E = \text{Cohen's } d = \frac{\mu_2 - \mu_1}{\sigma_{pooled}}$$

- For a single mean versus a null hypothesis:

$$E = \frac{\mu - H_0}{\sigma}$$

- In general, effect size is the relevant difference in the original scale divided by the standard deviation for that effect.

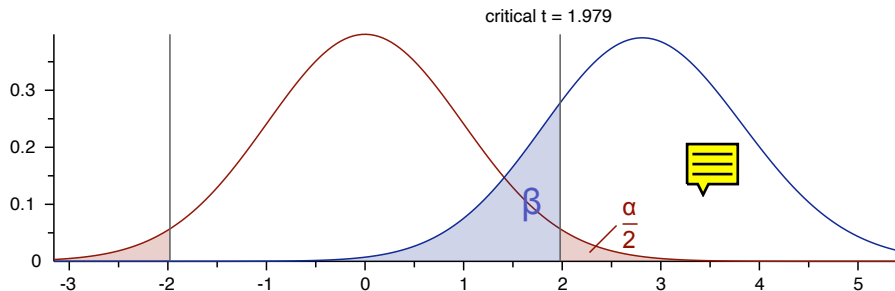
Outline

1. Preliminaries
2. Digging In
3. Various Examples
4. Summary

How Does Power Analysis Work?

- We assume the effect is true and put it at the center of the sampling distribution for the relevant sample statistic.
- As we know, the dispersion of the sampling distribution shrinks as n increases.
- For more power, a larger proportion of the area of this sampling distribution must sit beyond the rejection threshold for H_0 .
- e.g. if we want statistical power of .8, then 80% of the sampling distribution needs to be beyond the rejection threshold.
- Find the minimum sample size needed to make this happen.

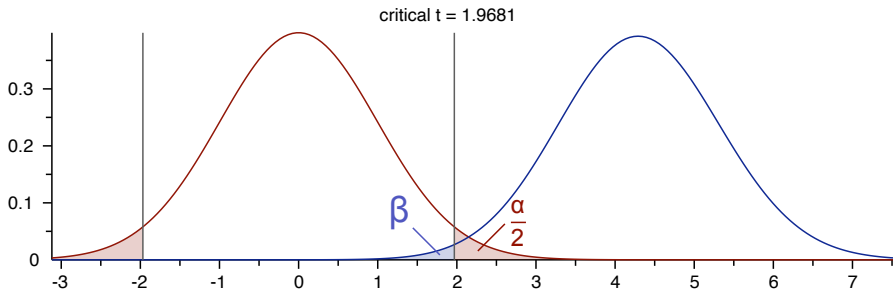
Illustration: Power of .80




The sampling distribution on the left represents $H_0 = 0$; the one on the right represents the assumed effect size as a t -stat.

Power is .8, since 80% of the area of the distribution around the effect lies above the threshold at which we would reject H_0 .

Illustration: Power of .99



This is the same scenario but statistical power is now .99. On the t-scale, the two distributions now are farther apart.

On the original scale of the  variable, they have shrunk. They look as wide as before because the t-scale converts into standard errors.

Using Software vs. Doing it By Hand

- It is quite easy to perform power analysis in Stata, R, or other applications, such as G*Power.
- One needs to understand how the process works to specify the commands properly, however.
- We will do one example by hand, deriving the equation needed for the minimum sample size.
- This is just for illustration. The math is tedious, so normally we will let software do it for us.

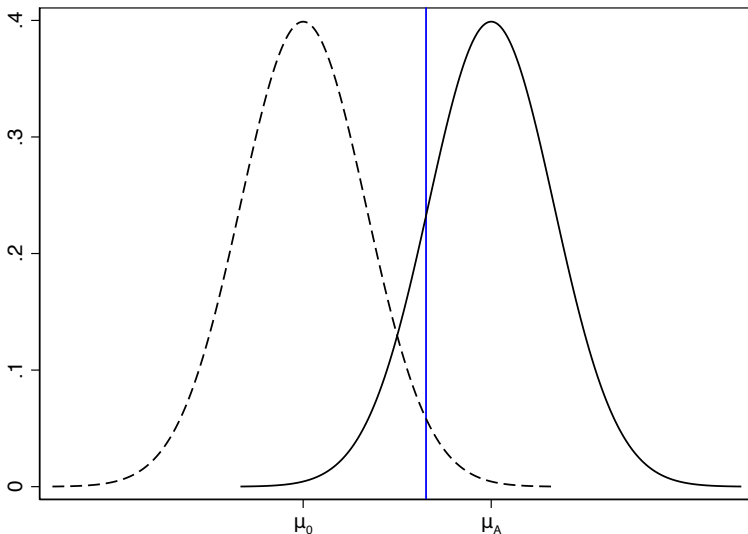
Logic Behind the Calculations

- Let's start with a one-sample test involving a mean.
- Suppose μ_A is the true mean but $H_0 = \mu_0$.
- Let's derive a formula to show when n is just large enough.
- Start with this scenario:

$$\mu_0 + t_\alpha \left(\frac{\sigma}{\sqrt{n}} \right) = \mu_A + t_\beta \left(\frac{\sigma}{\sqrt{n}} \right)$$

- The left side is the threshold at which we reject H_0 . The right side is the threshold for achieving the desired statistical power.
- Note: if $\mu_A > \mu_0$, then $t_\beta < 0$. If $\mu_A < \mu_0$, then $t_\alpha < 0$.

At the blue line: $\mu_0 + t_\alpha \left(\frac{\sigma}{\sqrt{n}} \right) = \mu_A + t_\beta \left(\frac{\sigma}{\sqrt{n}} \right)$.



Note: t_α is the critical value for rejecting H_0 ; t_β is negative here.

Derive the Formula for Minimum n

$$\begin{aligned}\mu_0 + t_\alpha \left(\frac{\sigma}{\sqrt{n}} \right) &= \mu_A + t_\beta \left(\frac{\sigma}{\sqrt{n}} \right) \\ t_\alpha \left(\frac{\sigma}{\sqrt{n}} \right) - t_\beta \left(\frac{\sigma}{\sqrt{n}} \right) &= \mu_A - \mu_0 \\ (t_\alpha - t_\beta) \left(\frac{\sigma}{\sqrt{n}} \right) &= \mu_A - \mu_0 \\ t_\alpha - t_\beta &= (\mu_A - \mu_0) \left(\frac{\sqrt{n}}{\sigma} \right) \\ t_\alpha - t_\beta &= \left(\frac{\mu_A - \mu_0}{\sigma} \right) \sqrt{n} \\ t_\alpha - t_\beta &= E\sqrt{n} \\ \left(\frac{t_\alpha - t_\beta}{E} \right)^2 &= n\end{aligned}$$

Now Use the Formula

$$n = \left(\frac{t_\alpha - t_\beta}{E} \right)^2 \text{ where } E = \frac{\mu_A - \mu_0}{\sigma}$$

- This formula will give us the minimum n needed for desired statistical power.
- We need to supply an estimate for σ .
- It's tricky because the t -statistics will change as n changes.
- Suggestion: start by assuming $t=z$. If n turns out to be small, then adjust the t -statistics for degrees freedom and recalculate.
- May have to repeat a few times until n stops changing

Example

$$n = \left(\frac{t_\alpha - t_\beta}{E} \right)^2 \text{ where } E = \frac{\mu_A - \mu_0}{\sigma}$$

- Suppose we expect that $\mu_A = 9$, while H_0 says $\mu_0 = 6$.
- Based on other data, we believe that σ is about 5.
- Thus, $E = \frac{3}{5} = .6$
- Let $t_\alpha = 1.96$.
- Suppose we want statistical power of .9. Find t_β such that 90% of the area lies above t_β . This is $t_\beta = -1.282$.

Example continued

$$\begin{aligned}n &= \left(\frac{t_{\alpha} - t_{\beta}}{E} \right)^2 \\n &= \left(\frac{1.96 - (-1.282)}{.6} \right)^2 \\n &= \left(\frac{3.242}{.6} \right)^2 \\n &= (5.40)^2 \\n &= 29.2\end{aligned}$$

- This suggests our minimum sample size is more like 30. But then the values of t would be different: 2.045 and -1.311.

Example continued with revised t -statistics

$$n = \left(\frac{t_\alpha - t_\beta}{E} \right)^2$$

$$n = \left(\frac{2.045 - (-1.311)}{.6} \right)^2$$

$$n = \left(\frac{3.356}{.6} \right)^2$$

$$n = (5.59)^2$$

$$n = 31.3$$

- Our minimum sample size is probably about 32.

Using Stata

- Stata can do these calculations with the `power` command.
- There are versions of this command for various types of power calculations: one sample means and proportions, two sample tests, correlation analysis, and many more.
- In this case, we will use the command for `onemean`.

```
power onemean 6 9, sd(5) power(.9)
```

- The null is listed first. We have to specify our estimate of σ and our desired power.

Stata Output: One Mean

```
. power onemean 6 9, sd(5) power(.9)
```

```
Performing iteration ...
```

```
Estimated sample size for a one-sample mean test  
t test
```

```
Ho:  $\mu = \mu_0$  versus Ha:  $\mu \neq \mu_0$ 
```

```
Study parameters:
```

alpha =	0.0500
power =	0.9000
delta =	0.6000
m0 =	6.0000
ma =	9.0000
sd =	5.0000

```
Estimated sample size:
```

N =	32
-----	-----------

Using R

- In R, one option is to use the `pwr` library, which you would first have to install.

```
install.packages("pwr")
```

- Then, specify the effect size as a formula, the power level, and the type of test.

```
> library(pwr)
> pwr.t.test(d=(9-6)/5, power=0.9, sig.level=0.05, type="one.sample", alternative="two.sided")
```

One-sample t test power calculation

```
      n = 31.17168
      d = 0.6
sig.level = 0.05
  power = 0.9
alternative = two.sided
```


Outline

1. Preliminaries
2. Digging In
3. Various Examples
4. Summary

One Sample Proportion

- We want sufficient power to identify when a sample proportion is meaningfully different from the null hypothesis.
- Suppose the airlines say 85% of flights arrive by the scheduled time.
- You would like to be able to claim that the true rate of on time arrival is 80% or less.
- For statistical power of .90, how large a sample of flights would you need to obtain?

```
power oneproportion .85 .80, alpha(.05) power(.9)
```

Stata Output: One Proportion

```
. power oneproportion .85 .80, alpha(.05) power(.9)
```

```
Performing iteration ...
```

```
Estimated sample size for a one-sample proportion test  
Score z test
```

```
Ho:  $p = p_0$  versus Ha:  $p \neq p_0$ 
```

```
Study parameters:
```

```
alpha =    0.0500  
power =    0.9000  
delta =   -0.0500  
p0 =      0.8500  
pa =      0.8000
```

```
Estimated sample size:
```

```
N =      589
```

R Output: One Proportion

```
> pwr.p.test(h=ES.h(p1=.80, p2=.85), sig.level=.05, power=.9, alternative="two.sided")  
  
proportion power calculation for binomial distribution (arcsine transformation)  
  
      h = 0.1318964  
      n = 603.9907  
sig.level = 0.05  
  power = 0.9  
alternative = two.sided
```

- Use the effect size function `ES.h(p1, p2)` to find the correct value for the first argument.
- The result is slightly different from Stata because here the binomial is used instead of the normal approximation.

Difference of Means Test

- With a difference of means test, we will have to make some additional educated guesses about parameters.
- The two samples each have a standard deviation. We need to estimate both.
- The samples may end up having different sizes:
 - ▶ If we will be doing random assignment, then the ratio of the two sample sizes will about 1 to 1.
 - ▶ If will be collecting observational data, we may have to guess the ratio of the sample sizes.

Difference of Means Scenario

- Suppose we plan to collect air quality samples from two cities and want to be able to detect a difference of 10 AQL.
- Let's imagine $\mu_1 = 150$ and $\mu_2 = 160$.
- We will sample both cities equally so the ratio of n_2 to n_1 is 1.
- From what we know based on the past, $\sigma_1 = 40$ and $\sigma_2 = 50$.

```
power twomeans 150 160, sd1(40) sd2(50)
```

or

```
power twomeans 150, diff(10) sd1(40) sd2(50)
```

Stata Output: Two Means

```
. power twomeans 150 160, sd1(40) sd2(50)
```

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: $m_2 = m_1$ versus Ha: $m_2 \neq m_1$

Study parameters:

alpha =	0.0500
power =	0.8000
delta =	10.0000
m1 =	150.0000
m2 =	160.0000
sd1 =	40.0000
sd2 =	50.0000

Estimated sample sizes:

N =	646
N per group =	323

R Output: Two Means

In R, we need to calculate the effect size (which is d below).

```
> m1 <- 150
> m2 <- 160
> sd1 <- 40
> sd2 <- 50
> sd.pooled <- sqrt((sd1^2 + sd2^2)/2)
> d <- (m2 - m1)/sd.pooled
> pwr.t.test(d=d, sig.level=0.05, power=0.80, type="two.sample")
```

Two-sample t test power calculation

```
      n = 322.7665
      d = 0.2208631
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

Variations of Difference of Means Power Analysis

- The default is to assume that both samples are the same size.
- We can add an option if we expect the ratio of n_2 to n_1 will be different.
- Suppose we think n_2 will be twice as large as n_1 :

```
power twomeans 150 160, sd1(40) sd2(50) nratio(2)
```

- Suppose we think n_2 will be half as large as n_1 :

```
power twomeans 150 160, sd1(40) sd2(50) nratio(.5)
```

```
. power twomeans 150, diff(10) sd1(40) sd2(50) nratio(2)
```

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: $m_2 = m_1$ versus Ha: $m_2 \neq m_1$

Study parameters:

alpha =	0.0500
power =	0.8000
delta =	10.0000
m1 =	150.0000
m2 =	160.0000
diff =	10.0000
sd1 =	40.0000
sd2 =	50.0000
N2/N1 =	2.0000

Estimated sample sizes:

N =	675
N1 =	225
N2 =	450

In R, the MESS library is needed.

```
> library(MESS)
> delta <- m2 - m1
> sd1 <- 40
> sd2 <- 50
> sd.pooled <- sqrt((.33*sd1^2 + .67*sd2^2))
> power_t_test(n=NULL, delta=delta, sd=sd.pooled, sig.level=.05,
+             type="two.sample", sd.ratio=5/4, ratio=2, power=.8)
```

Two-sample t test power calculation with unequal sample sizes and unequal variances

```
      n = 308.7889, 617.5778
delta = 10
      sd = 46.93613, 58.67016
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is vector of number in each group

Difference of Proportions Test

Earlier this semester we had the following data about a program to raise the rate of high school graduation:

Graduated?	Participant?		Total
	Yes	No	
Yes	83	102	185
No	10	19	29
Total	93	121	214
Proportion	.892	.843	.864

We failed to reject the null hypothesis. The program seems to make a difference, however. How large of a sample do we need to detect it?

Performing the Power Analysis

- Let's use the rates of graduation from the earlier study as our estimate of the proportions.
- If we call the students in the program sample 2, then the ratio of n_2 to n_1 is about .77.
- Let's say we want power of .8.
- This is all we need for the Stata command:

```
power twoproportions .843 .892, nratio(.77) power(.8)
```

or

```
power twoproportions .843, diff(.049) nratio(.77) power(.8)
```

```
. power twoproportions .843 .892, nratio(.77) power(.8)
```

Performing iteration ...

Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test

Ho: $p_2 = p_1$ versus Ha: $p_2 \neq p_1$

Study parameters:

alpha =	0.0500	
power =	0.8000	
delta =	0.0490	(difference)
p1 =	0.8430	
p2 =	0.8920	
N2/N1 =	0.7700	

Estimated sample sizes:

N =	1,540
N1 =	870
N2 =	670
N2/N1 =	0.7701

```
> library(MESS)
> power_prop_test(p1 = .843, p2 = .892, sig.level = .05, power = .8, ratio = .77)
```

Two-sample comparison of proportions power calculation with unequal sample sizes

```
      n = 869.2558, 669.3270
    p1 = 0.843
    p2 = 0.892
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is vector of number in each group

Outline

1. Preliminaries
2. Digging In
3. Various Examples
4. Summary

Other Power Tests and Software

- Stata has tests for dependent samples, correlation, bivariate regression, etc.
- There is also a very nice free application called G*Power for Mac and Windows that does power analysis.
- You can download it at the url below:

`http://www.psychologie.hhu.de/arbeitsgruppen/
allgemeine-psychologie-und-arbeitspsychologie/
gpower.html`

Extensions

- We just calculated minimum n using, effect size, α -level, and desired power.
- Any one of these things can be calculated as a function of the others (making some assumptions about other parameters).
- For example, we can do **sensitivity analysis**: determining what effect size we can detect given a particular n and level of power.
- But that's for another course . . .