Power and Design

EDP 619 Week 12

Dr. Abhik Roy

Welcome!

This is an absolutely minimalist overview of power analysis and the reasons it matters. If you want a deeper dive, consider enrolling in an advanced statistics course like Power and Sample Size for Multilevel and Longitudinal Study Designs for free. In the meantime, remember that you can explore power and its relationships interactively at Understanding Statistical Power and Significance Testing.

Additionally you may notice the following icons in the footnotes. These contain links to external sites that provide extra materials that may be of interest to you.

HTML
EXTRA CONTENT

PDF
PAPER MATERIAL

Rmd
GUIDED EXAMPLE

R SCRIPT

ONLINE VIDEO

Something Important

(Statistical) Power and its corresponding analysis are by far one of the most misunderstood concepts, in that a lot of people think they know what it is but simply miss the point.

Please note that you will need to recall some concepts covered in an introductory statistics course

Prerequisites

Before going ahead, make sure that you have a basic understanding of sampling and hypothesis testing. For a refresher, please take a look at both reviews on the next page

Review

If you would like a deeper dive on either area, please click on the icons below

Review

If you would like a deeper dive on either area, please click on the icons below

Sampling

Review

If you would like a deeper dive on either area, please click on the icons below

Sampling

Hypothesis Testing

(Statistical) Power

Definition

Definition

(Statistical) Power is the probability of avoiding a Type II error *aka* detecting an effect if it exists

Dodging False Negatives

Dodging False Negatives

Influences

Influences

> Effect size

Influences

> Effect size
>
> Sample size

Influences

- Effect size
- Sample size
- Significance level

Some Things You Should Be Aware of

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

This means if there are actual effects detected in 100 different studies with 80% power, then 80 out of 100 statistical tests will actually detect them

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

If you have an *underpowered* design, you may not be able detect a true effect

This means if there are actual effects detected in 100 different studies with 80% power, then 80 out of 100 statistical tests will actually detect them

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

This means if there are actual effects detected in 100 different studies with 80% power, then 80 out of 100 statistical tests will actually detect them

If you have an *underpowered* design, you may not be able detect a true effect

If you have an *overpowered* design, you may detect very small effects that are of no practical relevance[1]

The real problem isn't that *overpowered* experiments may reveal tiny significant effects, rather it is that many academic fields and popular science reporting standard in general emphasize "statistical significance" - *a meaningless term itself* - over effect sizes - aka practical significance. And if you're wondering, by "statistical significance" I mean the *p*-value which is generally garbage, should be used sparingly, and if you must use it, never report the findings by themselves because they are useless

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

This means if there are actual effects detected in 100 different studies with 80% power, then 80 out of 100 statistical tests will actually detect them

If you have an *underpowered* design, you may not be able detect a true effect

If you have an *overpowered* design, you may detect very small effects that are of no practical relevance[1]

We want the Type I error rate to be reasonably low which is why you often see a default of $\alpha = 0.05$

The real problem isn't that *overpowered* experiments may reveal tiny significant effects, rather it is that many academic fields and popular science reporting standard in general emphasize "statistical significance" - *a meaningless term itself* - over effect sizes - aka practical significance. And if you're wondering, by "statistical significance" I mean the *p*-value which is generally garbage, should be used sparingly, and if you must use it, never report the findings by themselves because they are useless

Some Things You Should Be Aware of

Power is usually set at 80%, but that is not a hard rule by any means

This means if there are actual effects detected in 100 different studies with 80% power, then 80 out of 100 statistical tests will actually detect them

If you have an *underpowered* design, you may not be able detect a true effect

If you have an *overpowered* design, you may detect very small effects that are of no practical relevance[1]

We want the Type I error rate to be reasonably low which is why you often see a default of $\alpha = 0.05$

We want power to be high enough to minimize a Type II error (β) so we reduce the chance of missing an effect

The real problem isn't that *overpowered* experiments may reveal tiny significant effects, rather it is that many academic fields and popular science reporting standard in general emphasize "statistical significance" - *a meaningless term itself* - over effect sizes - aka practical significance. And if you're wondering, by "statistical significance" I mean the *p*-value which is generally garbage, should be used sparingly, and if you must use it, never report the findings by themselves because they are useless

Using *(Statistical) Power*

A ***power analysis*** is the procedure that researchers can use to determine if the test contains enough power to make a reasonable conclusion. It can also be used to calculate the number of samples required to achieve a specified level of power

In a Nutshell

In a Nutshell

A ***power analysis*** is a calculation that can help you to determine a minimum sample size for your study

Four Components

Four Components

Expected effect size

standardized way of expressing the magnitude of the expected result of your study typically based on similar studies or a pilot study

Four Components

Expected effect size

standardized way of expressing the magnitude of the expected result of your study typically based on similar studies or a pilot study

Sample size

the minimum number of observations needed to observe an effect of a certain size with a given power level

Four Components

Expected effect size

standardized way of expressing the magnitude of the expected result of your study typically based on similar studies or a pilot study

Sample size

the minimum number of observations needed to observe an effect of a certain size with a given power level

Significance level (α)

the maximum risk of rejecting a true null hypothesis that you are willing to take - typically set at 5% or lower

Four Components

Expected effect size

standardized way of expressing the magnitude of the expected result of your study typically based on similar studies or a pilot study

Sample size

the minimum number of observations needed to observe an effect of a certain size with a given power level

Significance level (α)

the maximum risk of rejecting a true null hypothesis that you are willing to take - typically set at 5% or lower

(Statistical) Power

the likelihood that a test will detect an effect of a certain size if there is one - typically set at 80% or higher

Four Components

Expected effect size

standardized way of expressing the magnitude of the expected result of your study typically based on similar studies or a pilot study

Sample size

the minimum number of observations needed to observe an effect of a certain size with a given power level

Significance level (α)

the maximum risk of rejecting a true null hypothesis that you are willing to take - typically set at 5% or lower

(Statistical) Power

the likelihood that a test will detect an effect of a certain size if there is one - typically set at 80% or higher

If you have any of the three parameters above, then you can also calculate the fourth one

A Rundown of Power Analysis in R

Getting Ready

To follow along, please make sure to do the following

Getting Ready

To follow along, please make sure to do the following

1. Open up a blank `.R` script

Getting Ready

To follow along, please make sure to do the following

1. Open up a blank `.R` script
2. Run `install.packages("pwr", dependencies = TRUE)` in the Console

Getting Ready

To follow along, please make sure to do the following

1. Open up a blank `.R` script

2. Run `install.packages("pwr", dependencies = TRUE)` in the Console

3. Load up the `pwr` package by running

```r
library(pwr)
```

List of Commands

Copy CSV Excel PDF

Syntax	Description
`pwr.2p.test`	Two proportions (equal *n*)
`pwr.2p2n.test`	Two proportions (unequal *n*)
`pwr.anova.test`	Balanced one-way ANOVA
`pwr.chisq.test`	Chi-square test
`pwr.f2.test`	General linear model
`pwr.p.test`	Proportion (one sample)
`pwr.r.test`	Correlation
`pwr.t.test`	*t*-tests (one sample, two samples, paired)
`pwr.t2n.test`	*t*-test (two samples with unequal *n*)

Showing 1 to 9 of 9 entries

t-test

Equal Groups

```
pwr.t.test(n=, d=, sig.level=, power=, type=, alternative=)
```

Equal Groups

Survey Research Methods

```
pwr.t.test(n=, d=, sig.level=, power=, type=, alternative=)
```

Option	Description
`n`	Sample size
`d`	Power level
`sig.level`	Significance level (default of 0.05)
`type`	Choice of *t*-test - either `"two.sample"` (default), `"one.sample"`, or `"paired"`
`alternative`	Choice of direction - either `"two.sided"` (default), `"less"`, or `"greater"`

```r
# Equal Groups

pwr.t.test(d = 0.8,
           sig.level = 0.05,
           power = 0.8,
           type = "two.sample",
           alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 25.52458
##               d = 0.8
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Unequal Groups

```
pwr.t2n.test(n1=, n2=, d=, sig.level=, power=, type=, alternative=)
```

Unequal Groups

```
pwr.t2n.test(n1=, n2=, d=, sig.level=, power=, type=, alternative=)
```

Option	Description
`n1`	One sample size
`n2`	The other sample size
`d`	Power level
`sig.level`	Significance level (default of 0.05)
`type`	Choice of *t*-test - either `"two.sample"` (default), `"one.sample"`, or `"paired"`
`alternative`	Choice of direction - either `"two.sided"` (default), `"less"`, or `"greater"`

```r
# Unequal Groups

pwr.t2n.test(n1 = 28,
             n2 = 35,
             d = 0.5)
```

```
##
##      t test power calculation
##
##              n1 = 28
##              n2 = 35
##               d = 0.5
##       sig.level = 0.05
##           power = 0.4924588
##     alternative = two.sided
```

Survey Research Methods

ANOVA

ANOVA

```
pwr.anova.test(k=, n=, f=, sig.level=, power=)
```

ANOVA

```
pwr.anova.test(k=, n=, f=, sig.level=, power=)
```

Option	Description
`k`	Number of groups
`n`	Common sample size in each group
`f`	Effect size
`sig.level`	Significance level (default of 0.05)
`power`	Power level

```r
pwr.anova.test(k = 3,
               f = 0.25,
               sig.level = 0.05,
               power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 68.49707
##               f = 0.25
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

Correlations

Correlations

```
pwr.r.test(n=, r=, sig.level=, power=, alternative=)
```

Correlations

```
pwr.r.test(n=, r=, sig.level=, power=, alternative=)
```

Option	Description
`n`	Sample size
`r`	Effect size
`sig.level`	Significance level (default of 0.05)
`power`	Power level
`alternative`	Choice of direction - either `"two.sided"` (default), `"less"`, or `"greater"`

```r
pwr.r.test(r = 0.40,
           sig.level = 0.05,
           power = 0.80,
           alternative = "greater")
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 36.50995
##               r = 0.4
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```

Survey Research Methods

General Linear Models

General Linear Models

```
pwr.f2.test(u=, v=, f2=, sig.level=, power=)
```

General Linear Models

```
pwr.f2.test(u=, v=, f2=, sig.level=, power=)
```

Option	Description
`u`	Numerator of degrees of freedom (df)
`v`	Denominator of degrees of freedom (df)
`f2`	Effect size
`sig.level`	Significance level (default of 0.05)
`power`	Power level

```r
pwr.f2.test(u = 2,
            f2 = 0.3/(1 - 0.3),
            sig.level = 0.001,
            power = 0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 49.88971
##              f2 = 0.4285714
##       sig.level = 0.001
##           power = 0.8
```

Tests of Proportions

Equal Groups

```
pwr.2p.test(h=, n=, sig.level=, power=)
```

Equal Groups

```
pwr.2p.test(h=, n=, sig.level=, power=)
```

Option	Description
`h`	Effect size
`n`	Sample size
`sig.level`	Significance level (default of 0.05)
`power`	Power level

```r
# Equal Groups

pwr.2p.test(h = ES.h(p1 = 0.55, p2 = 0.50)
            sig.level = 0.05,
            power = 0.80)
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.1001674
##               n = 1564.529
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

Unequal Groups

```
pwr.2p2n.test(h =, n1 =, n2 =, sig.level=, power=)
```

Unequal Groups

```
pwr.2p2n.test(h =, n1 =, n2 =, sig.level=, power=)
```

Option	Description
`h`	Effect size
`n1`	One sample size
`n2`	The other sample size
`sig.level`	Significance level (default of 0.05)
`power`	Power level
`alternative`	Choice of direction - either `"two.sided"` (default), `"less"`, or `"greater"`

```r
# Unequal Groups

pwr.2p2n.test(h = 0.2,
              n1 = 763,
              power = 0.8,
              sig.level = 0.05,
              alternative = "greater")
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.2
##              n1 = 763
##              n2 = 193.8285
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: different sample sizes
```

Chi-square

Chi-square

```
pwr.chisq.test(w =, N =, df =, sig.level =, power = )
```

Chi-square

```
pwr.chisq.test(w =, N =, df =, sig.level =, power = )
```

Option	Description
`w`	Effect size
`N`	Total sample size
`df`	Degrees of freedom
`sig.level`	Significance level (default of 0.05)
`power`	Power level

```r
pwr.chisq.test(w = 0.1,
               power = 0.9,
               df = 1,
               sig.level = 0.01)
```

```
##
##      Chi squared power calculation
##
##               w = 0.1
##               N = 1487.939
##              df = 1
##       sig.level = 0.01
##           power = 0.9
##
## NOTE: N is the number of observations
```

Survey Research Methods

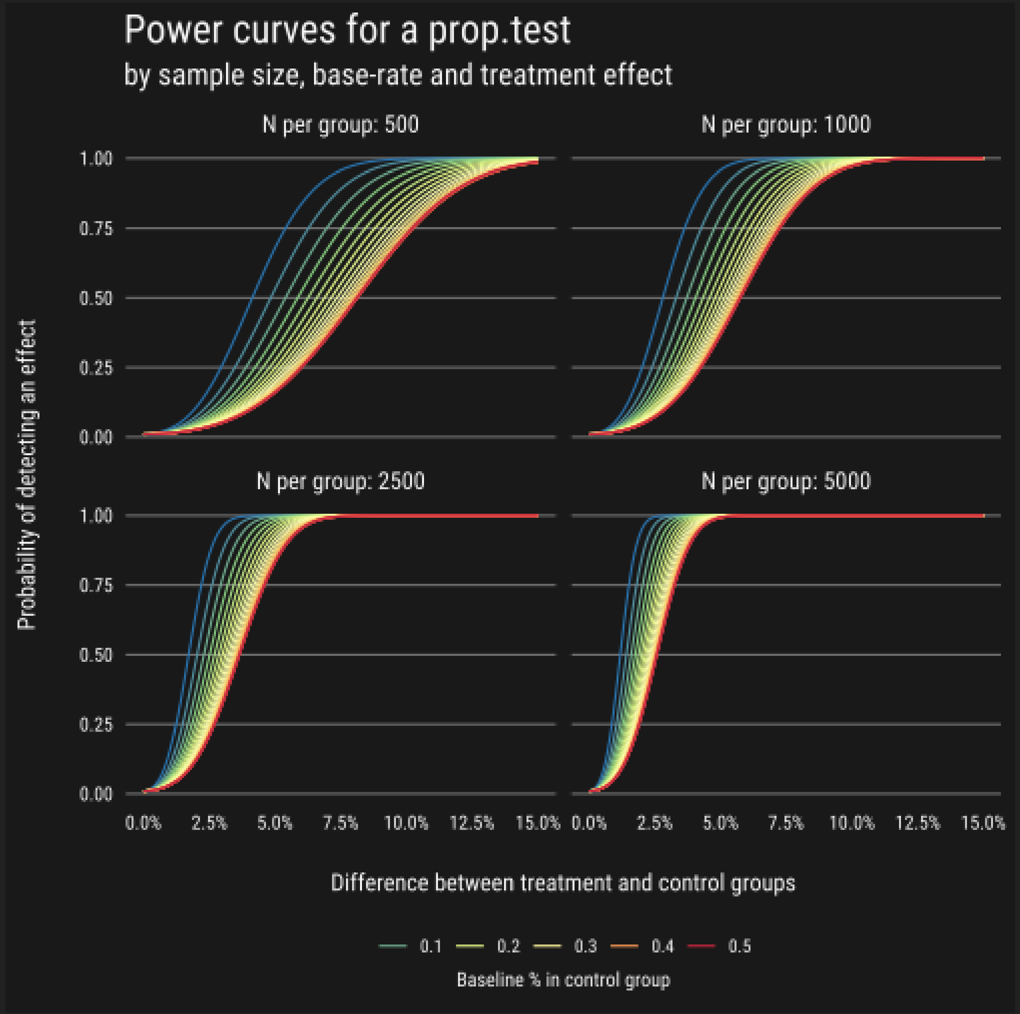
Power Analysis Plots

Power Analysis Plots

Power curves are line plots that show how the change in variables, such as effect size and sample size, impact the power of the statistical test

Unfortunately we do not have the bandwidth to cover power curves here, but an example and annotated script is provided on the next slide should you be interested

Example


Power curves for a prop.test
by sample size, base-rate and treatment effect
N per group: 500
N per group: 1000
N per group: 2500
N per group: 5000
Probability of detecting an effect
1.00
0.75
0.50
0.25
0.00
0.0%
2.5%
5.0%
7.5%
10.0%
12.5%
15.0%
Difference between treatment and control groups
0.1
0.2
0.3
0.4
0.5
Baseline % in control group

Note: Choosing a Starting Effect Size

Note: Choosing a Starting Effect Size

> Estimating an expected effect size is the most difficult parameter to determine

Note: Choosing a Starting Effect Size

> Estimating an expected effect size is the most difficult parameter to determine
>
> It requires that you have experience with the content and context of a study and any corresponding measures (e.g. using previous studies and results to calculate possible effect sizes)

Note: Choosing a Starting Effect Size

> Estimating an expected effect size is the most difficult parameter to determine
>
> It requires that you have experience with the content and context of a study and any corresponding measures (e.g. using previous studies and results to calculate possible effect sizes)
>
> If you don't have access to this information or if a study is completely novel, Cohen (1988) created some basic estimations and benchmarks. Much like other guidelines such as *p*-values (ugh), Cohen's Kappa, etc, these serve as a starting point and by no means should they be treated as static rules!

Note: Choosing a Starting Effect Size

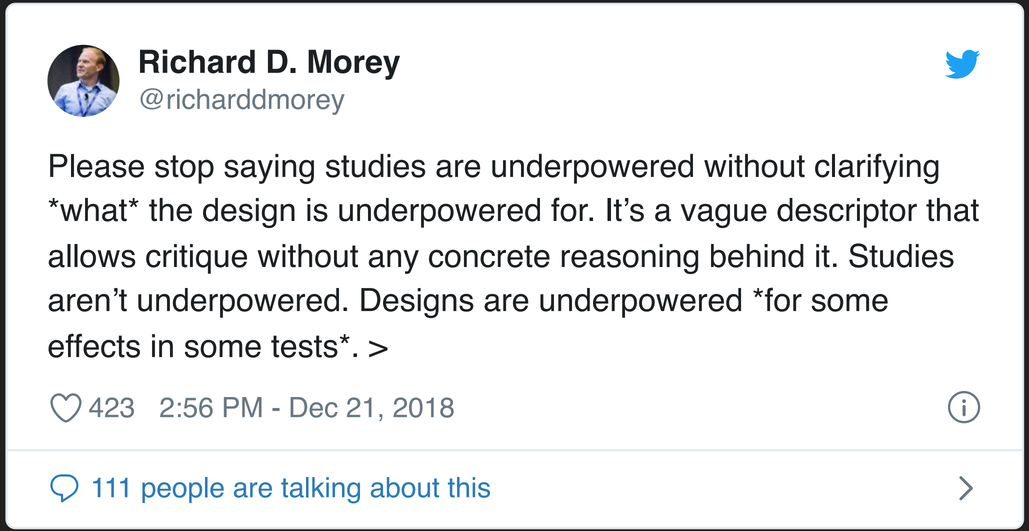
> Estimating an expected effect size is the most difficult parameter to determine
>
> It requires that you have experience with the content and context of a study and any corresponding measures (e.g. using previous studies and results to calculate possible effect sizes)
>
> If you don't have access to this information or if a study is completely novel, Cohen (1988) created some basic estimations and benchmarks. Much like other guidelines such as *p*-values (ugh), Cohen's Kappa, etc, these serve as a starting point and by no means should they be treated as static rules!

Approach	Variable	Small Effect	Medium Effect	Large Effect
t-test	`d`	0.20	0.50	0.80
ANOVA	`f`	0.10	0.25	0.40
General Linear Models	`f2`	0.02	0.15	0.35
Tests of Proportions	`h`	0.20	0.50	0.80
Chi-square	`w`	0.10	0.30	0.50

PDF

Terminology is Important!

How you describe power is important!



Terminology is Important!

How you describe power is important!

For more on this, click the graphic to the right

Additional Resources

If you would like a different view of power and its use, please click on the icons below

Additional Resources

Survey Research Methods

If you would like a different view of power and its use, please click on the icons below

(Statistical) Power

Additional Resources

If you would like a different view of power and its use, please click on the icons below

(Statistical) Power

Power Analysis

Thats it!

If you have any questions, please reach out

Thats it!

If you have any questions, please reach out