Power Analysis and Sample Size Calculation

Laura Saba

November 15, 2018

What this lecture covers

- Introduction to R
- Overview of power analyses and sample size calculations
- Power analyses for t-tests
- Power analyses for comparing proportions
- Power analyses for correlation
- Alternatives to R

Introduction to R

Online tutorials

- **Software Carpentry** - http://software-carpentry.org/lessons/
- **Data Carpentry** - http://www.datacarpentry.org/lessons/
- **My Software Carpentry Lesson** - https://github.com/LauraSaba/softwarecarpentry_R_2016-11-04
- **Try R - Code School** - http://tryr.codeschool.com/

R as a calculator

The R "console" is where you can interact with R directly. The $>$ sign is the R "prompt". It indicates that R is waiting for you to type something.

Let's start by subtracting a couple of numbers.

```
2016 - 1969
```

```
## [1] 47
```

R does the calculation and prints the result, and then you get the $>$ prompt again.

You can use R as a calculator in this way.

Need for Scripts

We can go along, typing directly into the R console. But there won't be an easy way to keep track of what we've done.

It's best to write R "scripts" (files with R code), and work from them.

Interacting with R

While you can type R commands directly at the > prompt in the R console, I recommend typing your commands into a script, which you'll save for later reference, and then executing the commands from there.

Start by typing the following into the R script.

```
# R intro


2016 - 1969
```

Save the file clicking on File–>Save, or by typing Ctrl + S (Windows).

Now place the cursor on the line with 2016 - 1969 and type Ctrl + Enter. The command will be copied to the R console and executed, and then the cursor will move to the next line.

You can also highlight a bunch of code and execute the block all at once with Ctrl + Enter.

Commenting

Use # signs to comment. Anything to the right of a # is ignored by R, meaning it won't be executed. Comments are a great way to describe what your code does within the code itself, so comment liberally in your R scripts.

Power analyses and sample size calculations

Most of what I am going to talk about today is copied from:

http://www.statmethods.net/stats/power.html

Power analysis overview

Power analysis is an important aspect of experimental design.

1. It allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence.
2. Conversely, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints. If the probability is unacceptably low, we would be wise to alter or abandon the experiment.

Values needed for calculations

The following 4 quantities have an intimate relationship:

1. sample size
2. effect size
3. significance level (probability of a Type I error given there is no effect)
4. power (1 - probability of a Type II error given an effect of a particular size)

Given any three, we can determine the fourth.

pwr package in R

For specialized calculations, R needs to install and load packages. There are thousands of packages for R, but today we are only interested in *pwr*.

To use an R package, you first need to install it:

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

This gets the library on your machine. You won't have to install the package again unless you update your version of R.

Loading the *pwr* package into R

Now that we have installed *pwr* it is on our computer, but it isn't usable until we load it using the 'library' function.

```
library(pwr)
```

Power analyses for two-sample t-tests

In an ideal situation, we are interested in the sample size needed to detect a clinically relevant effect size. In this case, we can specify significance level, power, and effect size.

The *pwr* package has a function called 'pwr.t.test' that we can use to calculate the sample size.

pwr.t.test

'pwr.t.test' has 6 input parameters:

1. n - number of observations per group
2. d - effect size
3. sig.level - significance level
4. power - power of test
5. type - type of tests ("two.sample" or "one.sample" or "paired")
6. alternative - alternative hypothesis ("two.sided" or "greater" or "less")

Effect size - two-sample t-test

The effect size in a two-sample t-test is:

$d = \frac{|\mu_1 - \mu_2|}{\sigma}$

where $\mu_1 =$ mean of group 1, $\mu_2 =$ mean of group 2, and $\sigma =$ common standard deviation

Cohen suggests that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Sample size example - two-sample t-test

You want to design an experiment to determine the difference in deoxycytidine triphosphate (dCTP) concentration between your control and treatment groups. How many samples do you need per group to detect a medium effect size (effect size = 0.5)?

n = ?
d = 0.5
sig.level = 0.05
power = 0.80
type = "two.sample"
alternative = "two.sided"

Sample size example - two-sample t-test (R code)

```
pwr.t.test(d=0.5,sig.level=0.05,power=0.80,
           type="two.sample",alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Effect size - two-sample t-test

An "medium" effect size sounds fairly arbitrary. What if we want to know the difference in number of femtomole per million cells that represents a "medium" effect size?

We first need to determine the within group standard deviation. We can get it from similar data that has already been published or from pilot data.

Let's say that we estimate $\sigma = 100$:

$d = \frac{|\mu_1 - \mu_2|}{\sigma}$
$0.5 = \frac{|\mu_1 - \mu_2|}{100}$
$|\mu_1 - \mu_2| = 0.5 \times 100$
$|\mu_1 - \mu_2| = 50$

Reporting sample size calculation - two-sample t-test

With 64 samples per group, a significance threshold of 0.05, and a within group standard deviation of 100 femtomoles per million cells, we have 80% power to detect a difference of 50 femtomoles between groups.

Detectable effect size - two-sample t-test

More often due to finanical/time/etc. constraints, we are limited to a specific sample size and want to know what kind of difference we can detect. In this case, we can specify significance level, power, and sample size.

Effect size example - two-sample t-test

You want to design an experiment to determine the difference in deoxycytidine triphosphate (dCTP) concentration between your control and treatment groups. You can only collect 10 samples per group. What is the detectable effect size at 80% power and a significance threshold of 0.05?

n = 10
d = ?
sig.level = 0.05
power = 0.80
type = "two.sample"
alternative = "two.sided"

Effect size example - two-sample t-test (R code)

```r
pwr.t.test(n=10,sig.level=0.05,power=0.80,
           type="two.sample",alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##               d = 1.324947
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Effect size - two-sample t-test

Our power analysis indicated that we can detect an effect size of d = 1.325, an extremely large effect size according to Cohen. What is the difference in number of femtomole per million cells that this effect size represents?

Let's say that we estimate $\sigma = 100$ again:

$d = \frac{|\mu_1 - \mu_2|}{\sigma}$
$1.325 = \frac{|\mu_1 - \mu_2|}{100}$
$|\mu_1 - \mu_2| = 1.325 \times 100$
$|\mu_1 - \mu_2| = 132.5$

Reporting sample size calculation - two-sample t-test (version 2)

With 10 samples per group, a significance threshold of 0.05, and a within group standard deviation of 100 femtomoles per million cells, we have 80% power to detect a difference of 132.5 femtomoles between groups.

Variations of 'pwr.t.test'

We can also do similar calculations for paired t-test using the 'type="paired" ' option.

Here the effect size is: $d = \frac{|\mu_d iff|}{\sigma_{diff}}$

And n is the number of pairs.

Paired t-test power analysis in R

```r
pwr.t.test(d=0.5,sig.level=0.05,power=0.80,
           type="paired",alternative="two.sided")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

t-tests with unequal samples per group

So far, we have only considered study designs when there are equal numbers of observations per group. We can also specify a study design for a two-sample t-test that has a different number of samples in each group.

t-tests with unequal samples per group (R code)

```r
pwr.t2n.test(n1=50, n2=10, sig.level=0.05,power=0.80)
```

```
##
##      t test power calculation
##
##              n1 = 50
##              n2 = 10
##               d = 0.9869393
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Power analyses for comparing proportions

Within the *pwr*, there is also a function, 'pwr.2p.test', to determine the power to detect a specific difference in proportions.

pwr.2p.test

'pwr.2p.test' has 5 input parameters:

1. n - number of observations per group
2. h - effect size
3. sig.level - significance level
4. power - power of test
5. alternative - alternative hypothesis ("two.sided" or "greater" or "less")

Effect size - comparing two proportions

The effect size when comparing two proportions is:

$h = 2arcsin\left(\sqrt{p_1}\right) - 2arcsin\left(\sqrt{p_2}\right)$

where $p_1 =$ proportion in group 1 and $p_2 =$ proportion in group 2

Cohen suggests that h values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

Sample size example - comparing two proportions

You want to design an experiment to determine the difference in risk of heart attack between your control and treatment groups. How many samples do you need per group to detect a small effect size (effect size = 0.2)?

n = ?
h = 0.2
sig.level = 0.05
power = 0.80
alternative = "two.sided"

Sample size example - comparing two proportions (R code)

```r
pwr.2p.test(h=0.2,sig.level=0.05,power=0.80,
            alternative="two.sided")
```

```
##
##      Difference of proportion power calculation for bino
##
##               h = 0.2
##               n = 392.443
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

Sample size - comparing two proportions (cont.)

What if you want to know the sample size needed to detect a reduction of risk from 0.20 in the control group to 0.10?

Let's say that you estimate $p_1 = 0.20$:

$h = 2arcsin\left(\sqrt{p_1}\right) - 2arcsin\left(\sqrt{p_2}\right)$

```
p_1 = 0.20
p_2 = 0.10
h = 2*asin(sqrt(p_1)) - 2*asin(sqrt(p_2))
h
```

```
## [1] 0.2837941
```

Sample size - comparing two proportions (cont.)

```r
pwr.2p.test(h=0.284,sig.level=0.05,power=0.80)
```

```
##
##      Difference of proportion power calculation for binc
##
##               h = 0.284
##               n = 194.6256
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

Reporting sample size calculation - comparing two proportions

With 194 samples per group and a significance threshold of 0.05, we have 80% power to detect a reduction of risk from 0.20 to 0.10.

Detectable effect size - comparing two proportions

To determine the detectable effect size instead, you can specify significance level, power, and sample size.

Effect size example - comparing two proportions

You want to design an experiment to determine the difference in risk of a heart attack between your control and treatment groups. You can only collect 100 samples per group. What is our detectable effect size at 80% power and a significance threshold of 0.05?

n = 100
h = ?
sig.level = 0.05
power = 0.80
alternative = "two.sided"

Effect size example - comparing two proportions (R code)

```
pwr.2p.test(n=100,sig.level=0.05,power=0.80)
```

```
##
##      Difference of proportion power calculation for binc
##
##               h = 0.3962062
##               n = 100
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

Converting effect size to experiment relevant data - comparing two proportions

Because the power of comparing two proportions is dependent on the 'rarity' of the event. We first need to estimate the risk of heart attack in the control population. We can get it from similar data that has already been published or from pilot data.

Let's say that we estimate $p_1 = 0.20$:

$$h = 2arcsin\left(\sqrt{p_1}\right) - 2arcsin\left(\sqrt{p_2}\right)$$
$$2arcsin\left(\sqrt{p_2}\right) = 2arcsin\left(\sqrt{p_1}\right) - h$$
$$p_2 = \left[sin\left(\frac{2arcsin\left(\sqrt{p_1}\right) - h}{2}\right)\right]^2$$

```
h=0.396
p_1=0.20
p_2 = (sin((2*asin(sqrt(p_1))-h)/2))^2
p_2
```

```
## [1] 0.06892421
```

Reporting sample size calculation - comparing two proportions (version 2)

With 100 samples per group and a significance threshold of 0.05, we have 80% power to detect a reduction of risk from 0.20 to 0.07.

Comparing proportions with unequal samples per group

So far, we have only considered study designs when there are equal number of observations per group. We can also specify a study design for comparing proportions that has a different number of sample in each group.

Comparing proportions with unequal samples per group (R code)

```
pwr.2p2n.test(n1=50, n2=10, sig.level=0.05,power=0.80)
```

```
##
##      difference of proportion power calculation for binc
##
##                h = 0.9704947
##               n1 = 50
##               n2 = 10
##        sig.level = 0.05
##            power = 0.8
##      alternative = two.sided
##
## NOTE: different sample sizes
```

Power analyses for correlation analyses

Within the *pwr*, there is also a function, 'pwr.r.test', to determine the power to detect a correlation between two continuous variables (Pearson Correlation).

pwr.r.test

'pwr.r.test' has 5 input parameters:

1. n - number of observations
2. r - correlation coefficient
3. sig.level - significance level
4. power - power of test
5. alternative - alternative hypothesis ("two.sided" or "greater" or "less")

Effect size - correlation

The effect size when examining the correlation between two variables is simply the correlation coefficient.

Cohen suggests that r values of 0.1, 0.3, and 0.5 represent small, medium, and large effect sizes respectively.

Sample size example - correlation

You want to design an experiment to determine the association between heart rate and BMI. How many samples do you need per group to detect a large effect size (effect size = 0.5)?

n = ?
r = 0.5
sig.level = 0.05
power = 0.80
alternative = "two.sided"

Sample size example - correlation (R code)

```
pwr.r.test(r=0.5,sig.level=0.05,power=0.80)
```

```
##
##      approximate correlation power calculation (arctangh
##
##               n = 28.24841
##               r = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Reporting sample size calculation - correlation

With 29 samples and a significance threshold of 0.05, we have 80% power to detect a correlation coefficient of 0.5.

Detectable effect size - correlation

To determine the detectable effect size instead, we can specify significance level, power, and sample size.

Effect size example - correlation

You want to design an experiment to determine the correlation between heart rate and BMI. We can only collect 50 samples. What is our detectable effect size at 80% power and a significance threshold of 0.05?

n = 50
r = ?
sig.level = 0.05
power = 0.80
alternative = “two.sided”

Effect size example - correlation (R code)

```
pwr.r.test(n=50,sig.level=0.05,power=0.80)
```

```
##
##      approximate correlation power calculation (arctangh
##
##               n = 50
##               r = 0.3843186
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Reporting sample size calculation - correlation (version 2)

With 50 samples and a significance threshold of 0.05, we have 80% power to detect a correlation of 0.38.

Alternative to R - G*Power

G*Power is free downloadable program for sample size and power calculations

- software and manual are at http://www.gpower.hhu.de/
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160.
- older software that may not be consistently maintained

Alternative to R - PASS

PASS is a program for sample size and power calculations available for a fee

- PASS is available at https://www.ncss.com/software/pass/
- Although it is expensive ($395/year) it is well documented and comes with user support

Alternative to R - Individual Websites

Individual Websites

- two-sample t-test and difference in proportions - https://clincalc.com/stats/samplesize.aspx
- correlation - http://www.sample-size.net/correlation-sample-size/

What did we learn

- For power analyses/sample size calculations, you need to know 3 of the following to estimate the fourth; 1) sample size, 2) effect size, 3) significance level, and 4) power.
- The *pwr* package in R can calculate any of the four for several different types of statistical tests and study designs.
- More details on other functions within *pwr* for other study designs can be found at http://www.statmethods.net/stats/power.html