In this note, we discuss the use of Cohen’s D for planning difference-of-mean experiments.

**Estimating sample size**

Let’s imagine you are testing a new weight loss program and comparing it so some existing weight loss regimen. You want to run an experiment to determine if the new program is more effective than the old one. You’ll put a control group on the old plan, and a treatment group on the new plan, and after three months, you’ll measure how much weight the subjects lost, and see which plan does better on average.

The question is: how many subjects do you need to run a good experiment?

In order to answer that question, you generally need to decide several things:

* **The probability of a false positive that you are willing to tolerate.** This is the probability of declaring that the new weight loss plan is better, when it’s actually not. Also known as the *significance level threshold*, or “p-value threshold” of the experiment.
* **The probability of detecting a true positive that you want to achieve.** This is the probability of correctly detecting that the new weight loss plan is better than the original, when it actually is. Also known as the *power* of the experiment. You can also think of it as one minus the probability of a false negative.
* **How much of a difference you are willing to call a “real” difference.** If people on the new plan lose on average five more pounds every three months than they did on old plan, then the new plan is an improvement over the old one. On the other hand, if the average weight loss over three months on the two plans differs by less than a pound, you may not want to call that a “real” difference; you may want to consider the two plans “about the same.” The difference in average weight loss between the two plans is also known as the *effect size* observed in the experiment.

Let’s say that you are willing to tolerate a probability of either false posititve or false negative of 1%. This means that you want a p-value threshold of 0.01, and a power of 0.99.

p\_value <- 0.01

power <- 0.99

You also decide that you will call the new program “better” if the average weight loss at the end of the trial is two pounds or more greater on the new program than on the old.

**Cohen’s D: Normalized effect size**

Of course, not everyone on a program is going to lose exactly the same amount of weight; in fact, weight loss generally varies widely. Let’s say you know from experience that on average, people on the old program lose about 10 pounds after three months, and the standard deviation on the weight loss is about five pounds. This relatively wide standard deviation will make it harder to detect an average difference of two pounds than if the standard deviation were (miraculously) about one pound:

*Cohen’s d* is a measure of effect size for the difference of two means that takes the variance of the population into account. It’s defined as

d = | μ1 – μ2 | / σpooled

where σpooled is the pooled standard deviation over both cohorts.

σpooled = √( σ12 + σ22 )

Note that this formula assumes both cohorts are the same size.

The use of Cohen’s *d* for experimental design also assumes that the true standard deviation of the two treatment groups is about the same; only the mean differs.

Cohen suggested as a rule of thumb that *d* = 0.2 is a small effect, *d* = 0.5 is medium, and *d* = 0.8 is large ([Wikipedia has a more detailed rule-of-thumb table](https://en.wikipedia.org/wiki/Effect_size#Cohen's_d)). Here’s what those values of *d* (plus one more) look like:

**Cohen’s D and physical effect sizes**

But of course, we’d like to map *d* to meaningful physical units. If we assume that subjects on the new program will have about the same variance in weight loss as people on the old program, we can estimate the minimum *d* that we’d like to detect.

# in pounds

sigma <- 5

delta\_mu <- 2

(d\_minimum <- delta\_mu/sigma)

## [1] 0.4

Now we can estimate how many subjects we need per group, using the pwr.t.test() function from the pwr package in R. This function has four primary arguments:

* number of subjects per cohort,
* effect size (as *d*)
* significance level
* power

Given any three (and the fourth as NULL), pwr.t.test() estimates the value of the fourth. We’ll do a two-sided, two-sample test (compare two populations and check if the means are different).

library(pwr)

n\_per\_cohort <- NULL # what we want to know

effect\_size <- d\_minimum

p\_value <- 0.01

power <- 0.99

type <- "two.sample" # default value

alternative <- "two.sided"# default value

## given any three you can estimate the fourth

(estimate <- pwr.t.test(n = n\_per\_cohort,

d = effect\_size,

sig.level = p\_value,

power = power,

type = type,

alternative = alternative))

##

## Two-sample t test power calculation

##

## n = 302.0564

## d = 0.4

## sig.level = 0.01

## power = 0.99

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

We estimate that you will need about 303 subjects in each cohort (or 606 subjects total) to reliably detect a difference of two pounds average weight loss between the two groups of subjects (where “reliably” means 1% chance of a false positive and 1% chance of a false negative).

If this seems like too large an experiment, then you may need to relax your error bounds, or test for a larger effect size; whichever is more appropriate to your situation. If you are willing to reduce the power of your test (but leave the significance level where it is), plotting the output of pwr.t.test will give you a helpful power analysis graph.

# you could just plot it, but I want the data

eplot <- plot(estimate)

powanal <- eplot$data

# a function that returns the approx sample size for a given power

panal\_fun <- approxfun(powanal$power, powanal$sample\_sizes)

powers <- c(0.9, 0.95)

nframe <- data.frame(power=powers,

sample\_sizes=ceiling(panal\_fun(powers)))

nframe <- transform(nframe,

label = paste("approx n =", sample\_sizes))

library(ggplot2)

ggplot(powanal, aes(x=sample\_sizes, y=power)) +

geom\_point() + geom\_line(size = 0.1, color="red") +

geom\_hline(yintercept = nframe$power, linetype=3, color="darkblue") +

geom\_vline(xintercept = nframe$sample\_sizes, linetype=3, color="darkblue") +

geom\_text(data=nframe, aes(x = 300, y = power, label=label)) +

ggtitle("Two-sample t test power calculation",

subtitle = "n is sample size per group")

If you want to detect an effect size of at least two pounds to a significance of 0.01 and a power of 0.95, you need 226 subjects per group, or 452 subjects total. If you can settle for a test with a power of 0.9, then you only need 188 subjects per group, or 376 subjects total.

**Estimating minimum detectable effect size for a given sample size**

What often happens in real life is that the experimenter only has access to as many subjects as they can gather; they then hope that the set is large enough to detect any usedful difference. You can use pwr.t.test() to estimate the efficacy of your experiment in this situation, as well.

Suppose you are only able to gather 200 subjects (100 subjects for each diet plan). How big a difference can your experiment detect?

n <- 100

(estimate <- pwr.t.test(n = n,

d = NULL,

sig.level = p\_value,

power = power,

type = type,

alternative = alternative))

##

## Two-sample t test power calculation

##

## n = 100

## d = 0.6991277

## sig.level = 0.01

## power = 0.99

## alternative = two.sided

##

## NOTE: n is number in \*each\* group

# how large an effect in pounds?

(delta\_mu <- sigma \* estimate$d)

## [1] 3.495639

An experiment of this size, with a desired significance of 0.01 and a desired power of 0.99, can only reliably detect an effect size of about 3.5 pounds or larger. Lowering the test power to 0.8 will lower the minimum detectable effect size to about 2.5 pounds; this means that even if the new diet truly improves average weight loss by about 2.5 pounds over three months, your experiment will have a 20% chance of failing to detect it.

**Minimum detectable effect size as a function of experiment size**

Assuming that you want to keep your error bounds as they are, and you can gather a suitable number of subjects at will, what is the minimum detectable effect size for a given sample size? You can estimate that using pwr.t.test(), too.

sample\_sizes <- seq(from = 50, to = 500, by = 50)

# a function to get the effect size as d

get\_effect\_size <- function(n, sig\_level = 0.01, power = 0.99) {

test\_results <- pwr.t.test(n, NULL, sig\_level, power)

test\_results$d

}

# estimate d for a range of sample sizes

dvec = vapply(sample\_sizes,

function(n) {get\_effect\_size(n)},

numeric(1))

# convert dvec into approximate difference in pounds

# assuming both populations have the same sigma

sigma <- 5

diff\_frame <- data.frame(n = sample\_sizes,

delta\_lbs = dvec \* sigma)

# what's the approximate sample size (per cohort) to detect a difference of 3 pounds?

(n\_3 <- ceiling(approx(diff\_frame$delta\_lbs, diff\_frame$n, 3)$y))

## [1] 139

nframe <- data.frame(n = n\_3,

delta\_lbs = 3,

label = paste("approx n =", n\_3))

ggplot(diff\_frame, aes(x=n, y=delta\_lbs)) +

geom\_point() + geom\_line(size = 0.1, color="red") +

geom\_hline(yintercept = 3, color="darkblue", linetype=3) +

geom\_vline(xintercept = n\_3, color="darkblue", linetype=3) +

geom\_text(data = nframe, aes(y = 1.5, label=label)) +

ylab("Minimum detectable effect size (pounds)") +

ggtitle("Estimated minimum detectable effect size (pounds)")

To detect an average weight loss difference of three pounds to a significance of 0.01 and a power of 0.99, you will need about 139 subjects per group, or 278 total.

A general rule of thumb is that for a given set of error bounds, if you you want to halve the effect size you want to measure (go from detecting a difference of three pounds to 1.5 pounds), you need four times the data (from 139 subjects per group to 556 subjects per group).

This analysis assumes that both populations have about the same standard deviation of 5 pounds. If you suspect that the new weight loss program might have higher variance than the old one, then you should assume a higher pooled standard deviation, which means a smaller *d*, which means you will need more subjects.

**Testing differences in rates**

For testing the difference in rates of two processes, use Cohens’ *h*. The function ES.h calculates the appropriate *h* for a desired difference in rates, and the functions pwr.p.test(), pwr.2p.test(), and pwr.2p2n.test() help you estimate appropriate experiment sizes in different situations.

**Conclusion**

Our overall point is that proper experiments need to have stated goals and documented plans before being executed. Power calculators help you properly design the experiment size.