Power analysis

When we do what's called a power analysis, we need to be aware of several different quantities that all affect the power of a test. But before we dig into the details, it might be best to remember what power is:

$$\text{Power} = 1 - \beta = Pr\{reject\ H_0\ if\ H_0\ is\ false\}$$

When we discussed power earlier, we never talked about calculating *power*. And in one sense, it's just a little bit silly because to calculate power we have to make guesses as to the value of the parameters in the alternative hypothesis. But it can be a very useful tool in designing experiments.

1 The basic idea

Let's start off with a simple example. We will once again assume that we know the true value of σ. This isn't necessary, but it makes the sample calculations a lot simpler, and all we really want to do is to *understand* the idea. Once we do, we can let R take over as usual.

We'll use the deer hind/fore foot example from our discussion on paired tests. You may remember that for this example we used $\bar{d}$ instead of $\bar{y}$ as we were interested in looking at the differences. We will use the calculated value of s_d for σ to give us:

$$s_d = \sigma = 0.97$$

Now let's consider the following hypotheses (using $\alpha = 0.05$):

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

We need to consider for what values of $\bar{d}$ we would reject. It turns out we can calculate this easily if we assume we know σ:

To get the smallest value for $\bar{d}$ for which we would reject we do $\mu - z_{0.025}\frac{\sigma}{\sqrt{n}}$

To get the largest value for $\bar{d}$ for which we would reject we do $\mu + z_{0.025}\frac{\sigma}{\sqrt{n}}$

This would give us:

For the smallest value: $0 - 1.96\frac{0.97}{\sqrt{8}} = -0.672$

For the largest value: $0 + 1.96\frac{0.97}{\sqrt{8}} = 0.672$

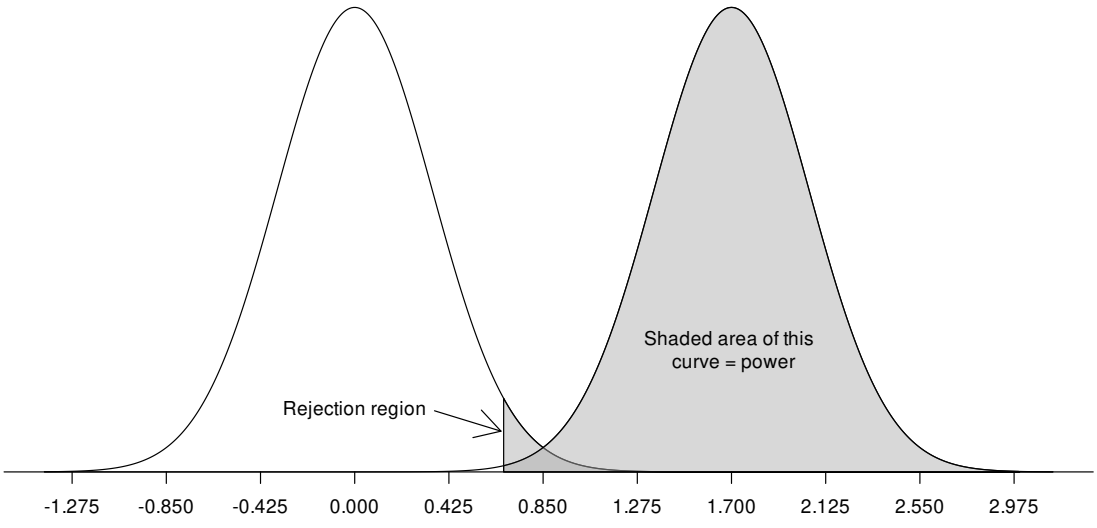
In other words, if $\bar{d} \leq -0.672$ or $\bar{d} \geq 0.672$ we would reject.
So how does this help us in calculating the power of our test? Well, suppose the alternative hypothesis is true and μ is really equal to 1.7 instead of 0. Our first normal curve (the one assuming H_0 is true) is centered at 0. If H_1 is true and we somehow knew that $\mu = 1.7$ the normal curve for H_1 would be centered at 1.7.

So let's draw both normal curves on the same axis (that's important). The question is, *how much of the curve for $\mu = 1.7$ is above the cut-off of 0.672*? But we can calculate this!

If $\mu = 1.7, \sigma = 0.97,$ and $n = 9$, what is $\Pr\{\bar{d} > 0.672\}$?

$$z = \frac{0.672 - 1.7}{\frac{0.97}{\sqrt{8}}} = -3.00$$

And from this we get $\Pr\{Z > -3\} = 0.9987$. In other words, the *power* of our test is 0.9987. That's pretty good - it means if the true value for μ really is 1.7 we would almost certainly reject.



Two normals centered at μ = 0, and μ = 1.7, both with σ = 0.97/√8

In order to do calculations of this type, we actually needed to know three of the following four quantities:

1. The value of α.
2. The sample size (n).
3. The alternate value of μ (i.e., the value of μ if H_1 is true).
4. The power $(1 - \beta)$.

We used the first three to calculate the fourth (power). But we can actually calculate *any* of these quantities if we know the values of the other three.

We also need to mention that item (3.) is often called the *effect* size. This describes the observed effect that we see. For example, the difference between our $H_0 : \mu = 0$ and our assumed value of μ under H_1 is 1.7 (using the absolute value):

$$\mu_{H_0} - \mu_{H_1} = 0 - 1.7 = -1.7$$

A very common preliminary item when in a study might be to determine what sample size we need to detect a given effect size at desired power. For example, we might want to be able to discover an effect size of 0.67, with a power of 0.85 using $\alpha = 0.05$. As mentioned, if we have three of the items in our list above, we can calculate the fourth.

It should also be mentioned that the effect size is usually standardized by dividing by the (in this case common) standard deviation. So our standardized effect size would be:

$$\text{Effect size} = \frac{0 - 1.7}{.97} = 1.75$$

An obvious question is "what is a good effect size". The obvious answer is that it depends on what you're trying to do and what the science behind your analysis is. There are some general guidelines about what makes a good effect size, but it really does depend on the specific situation. Cohen's d is one of the most commonly used measures of effect size and is a slight variation of what is given above (using estimates instead of parameters - which actually makes more sense):

$$d = \frac{\bar{y_1} - \bar{y_2}}{s}$$

Cohen goes on to define the following effect sizes:

Cohen's d	Effect size
Small	0.2
Medium	0.5
Large	≥ 0.8

But as mentioned, this really depends on the science. Cohen, for example, was a behavioral scientist and worked mostly in psychology.

So where does that leave us? We will, as usual, resort to R to do our calculations.

2 Power analysis in R

To do power analysis in R we need to use the `pwr` package. This will calculate any of the four quantities given above if you give it the other three. Assuming it's installed and loaded (e.g., `library(pwr)`) here is a simple example using our above analysis, *but* now using the (correct) t-test approach instead of what we did above which was based on σ instead of s:

```
library(pwr)

pwr.t.test(n = 8, d = 1.75, sig.level = 0.05, type = "one.sample")
```

Most of the options should be obvious, but let's define them anyway:

> `n =` is the sample size (should be obvious for paired or one sample tests; for two sample tests, `n` is the sample size in *each* sample assuming $n_1 = n_2$. If you have unequal sample sizes for a two sample test use the `pwr.t2n.test` instead.
>
> `d =` is the effect size.
>
> `sig.level =` is our desired value of α.
>
> `type =` is the type of t.test. Valid options are `"one.sample", "two.sample", "paired"`. The function defaults to a two sample test.

There are two more options not visible above:

> `p =` is the power you want; we didn't use this because we are interested in calculating the power. Remember this function can calculate *any* of the four quantities (sample size, power, effect size, significance level) given the other three.
>
> `alternative =` tells us if the alternative hypothesis should be directional; options are `"two.sided", "less", "greater"`. The function defaults to a two sided alternative.

In any case the above command gives results in R:

```
 Two-sample t test power calculation

              n = 8
              d = 1.75
      sig.level = 0.05
          power = 0.9874334
    alternative = two.sided

NOTE: n is number in *each* group
```

From this we can see that the power is 0.9874334. The difference in power here compared to what we calculated above is that R is (correctly) using the t distribution (with the appropriate *df*) instead of the normal distribution. If you remember, the t distribution has fatter tails, particularly at low n, so there will be more overlap between the distributions and the power will be lower.

The `pwr` package also has commands for tests of proportion, ANOVA (one way, equal n), χ^2, correlation as well as for a general linear model approach. For all of these, the idea is the same: give R three of the four quantities, and it will calculate the missing quantity. For more details you might look up Power Analysis on the "Quick-R" web page:

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

(This is a pretty decent collection of web pages for simple explanations on using R).