The power of a statistical test informs one how likely it is that the null hypothesis will be accepted even though the null is not true.

Since our alternatives don't fully specify the alternative, the power is typically defined in terms of some "effect size." That is, if there is a known value for the alternative (say $\mu = 2$ for example) then the power is 1 minus the probability the null will be "accepted" even though it is not actually true.

We have a demo to help guide us. Type in the following command into your R session:

```
> source("http://wiener.math.csi.cuny.edu/verzani/classes/MTH214/ex-power.R")
```

This is a slightly modified version of the power demonstration found under the plot menu of `pmg`.

Let's keep this table straight in our mind

difference = observed-expected	small	large
z-scores (observed)	close to 0	not close to 0
significance	not statistically significant	statistically significant
H_0	Accept	Reject

Table 1: Different ways to say the same thing

The demonstration draws two densities. The bottom one illustrates the sampling distribution of $Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ assuming H_0 is true, although since there is no scale, you can think of this as the distribution of $\bar{x}$ itself. The area shaded in blue is equal to α, the significance level. Basically, **if** the observed value of Z ($\bar{x}$) were to land in this area, then the p-value would be less than α and otherwise, that is if the observed value were to land in the unshaded area, more than α.

Then, if we have our test statistic's value in the unshaded area then the p-value is bigger than α and so we would "accept" H_0.

Suppose in fact that the alternative was true. Then this would be a problem – a type-II error:

	Accept H_0	Reject H_0
H_0 is true	Good	Type-I error
H_0 is not true	Type-II error	Good

The probability of a Type-I error is exactly α. What is the probability of a type II error? It is 1 minus the power. In the picture this is $1 - 0.13 = 0.87$.

In terms of the picture, the red-shaded area for the alternative is correpsonds to the observed values that produces p-values that lead us to say "accept." So the red-shaded area is this probability. We would like the type-II error probability to be small (or the power large). How to get this?

Look at the top graphic drawing the alternative. This shows the sampling distribution of Z (or $\bar{x}$) *if* the alternative is true. The alternative is specified by a mean, and the demo uses the difference in means divided by the standard deviation to represent this. This is known as the **effect size**. The larger the effect, the easier it should be to see.

The default for the demo is 0.5 with a sample of size 1. That is, if σ is 1 and the true mean is 0.5 and there is a single observation from the population, then the power to reject H_0 is only 0.13. In otherwords, there is an 87% chance that we would falsely accept the null hypothesis.

Why? The answer is always the variability – We are trying to see if the difference between an observed and expected is large **compared** to the variability and when there is only one data point the variability is still large. (The variability related to $\sigma/\sqrt{n}$ or just σ when $n = 1$ as we have here.)

1. Even with just one sample, we can detect differences provided the effect is large enough. We just saw that an effect size of 0.5 is too small to reliably detect. By trial and error, change the effect size until the power is 0.8. What value for the effect size do you get.

Of course, we may not be able to just change our effect size. In fact, this is usually determined ahead of time as part of a study. So how can we change the variability? Since it is related to $\sigma/\sqrt{n}$ we simply change n – that is we take a bigger sample.

One mantra of this class is "the bigger the sample, the less variability." Go ahead say that 20 times.

1. Really, 20 times. You can mumble. Then write a 12 page report on how you felt. (Just checking if you are reading along here.)

Okay, how big a sample is required to detect an effect of size 1? with a power of 0.80? To investigate, we change the effect size to 1 and *carefully* adjust the slider for n. Around $n = 6$ we get a power of 0.79 – good enough for government work. To do better, we can actually crank out the probabilities – they use the normal distribution after all.

Cohen has several ranges for effect sizes. A *moderate* effect has an effect size of 0.5 and a *large* effect one of 0.8.

1. How large a sample is needed to detect a *large* effect with a power of 0.80?
2. How large a sample is needed to detect a *moderate* effect with a power of 0.90?
3. How large a sample is needed to detect a *small* effect (say a 0.2 effect size) with a power of 0.90?

The demo allows you to switch from a two-sided test to a one-sided test. What happens when this is done?

1. Start with an effect size of 0.5, a sample size of $n = 1$ and a "greater" test. Changing the test to "less" will make no sense here. Why?
2. If you change the test to a two-sided test does the power go up or down? Can you say why?

The power of a test depends on the choice of significance level, α.

1. Explain why making alpha bigger increases the power of the test.