Hypothesis testing, significance and power

In many research situations, the researcher is trying to collect evidence to decide between the "null hypothesis" that predictor variables do not have an effect on a response of interest and the "alternative hypothesis" that the predictor variables do have an effect on the response variable. Carrying out a study with the goal of choosing between two competing hypotheses is called "hypothesis testing".

Examples:

1. A sociology researcher is interested in how participation in after school activities effects school success. He may be interesting in collecting evidence to decide between the hypotheses:

   Null (H_0): there is no difference in school performance between students who participate in after school activities and those who do not.

   Alternative (Ha): there is a difference in school performance between students who participate in after school activities and those who do not.

2. A chemist is interested in if the pH of a solution affects the rate at which two solutes react. She may phrase this question in terms of the following competing hypotheses:

   Null (H_0): pH of the solution does not affect the reaction rate of the solutes.

   Alternative (H_a): pH of the solution affects the reaction rate of the solutes.

When a researcher performs a data analysis to decide between the null and alternative hypothesis he or she has formulated, there are four possible outcomes, which may be organised in the following table:

	The null hypothesis is true	The alternative hypothesis is true
The researcher accepts the null hypothesis	Correct decision	Incorrect decision (Type II error)
The researcher accepts the alternative hypothesis	Incorrect decision (Type I error)	Correct decision

If the null hypothesis is true and the data analysis leads the researcher to accept it, then the researcher has made a correct decision. Similarly, if the alternative hypothesis is true and the researcher is led to accept it, he or she has made a correct decision. However, there are two possible mistakes the researcher can make. If the null hypothesis is true but the researcher rejects it, this is called a Type I error. If the alternative hypothesis is true but the researcher rejects it, this is called a Type II error.

It is possible given information about a study to calculate the risks of making Type I and Type II errors. The probability of making a Type I error is often denoted by the Greek letter α, and is related to the significance level of a study by the formula

$$\text{Significance level} = 100(1-\alpha)\%$$

So if $\alpha=0.05$, this corresponds to a 95% significance level. This means that if the null hypothesis is rejected when the p-value calculated in the statistical test carried out on the data is $p<0.05$, then there is only a 5% chance of having made a Type I error. That is, if you carried out the study 100 times, you would expect to incorrectly reject the null hypothesis about 5 times.

The probability of making a Type II error is often denoted by the Greek letter β, and is related to the power of a statistical test by the formula

$$\text{Statistical power} = 100(1-\beta)\%$$

So $\beta=0.2$ corresponds to an 80% statistical power. The statistical power of a test depends on how far from true the null hypothesis is, which is called the effect size. For example:

1. If the average end of year marks for students participating in after school activities is 10% higher than the average end of year marks for students who do not participate in after school activities, then the effect size is 10%.
2. If the reaction rate of the solutes is 1.3 moles/liter/minute faster for each decrease by 1 unit in pH of the solution, then the effect size is 1.3 moles/liter/minute.

A statistical power of 80% against an effect size ε at a significance level α means that if the study were carried out 100 times in a setting where the effect size is at least ε, and if you accept the null hypothesis if your calculated p-value is less than α, then you would expect to incorrectly accept the null hypothesis (and reject the alternative hypothesis) about 20 times. Alternatively, if you obtain a p-value greater than α from an experiment with 80% power against effect size ε, it is likely that the effect size is less than ε.