More about hypothesis testing

- **The experimenter's hypothesis**
  There is a tendency to want to identify the experimenter's claim as the null hypothesis. This is rarely correct: the null hypothesis is the one that the experimenter wants to rule against by marshalling evidence from the data!

- **Correctly interpreting the *P*-value**
  The *P*-value is *not* the probability that the null hypothesis is true. It is a conditional probability that, *if* the null hypothesis us true, the test statistic would be as extreme as it is observed to be.

- **Alpha level**
  In some hypothesis testing, we set a *predetermined* level, denoted α, below which the *P*-value from the test is deemed to be statistically significant: if the *P*-value is $\leq \alpha$, we reject H_0; if the *P*-value is $> \alpha$, we do not reject H_0. Standard values of α are 0.10, 0.05, 0.01, and sometimes even smaller values.

- **Statistical significance is not practical significance.**
  A test can offer statistical significant evidence to conclude that the value of a parameter is not equal to some value, but the actual difference between the true and hypothesized value could be so small as to be practically *in*significant. Look also at a confidence interval estimate for the parameter to help decide this.

- **Confidence intervals and hypothesis tests**
  Confidence intervals and hypothesis tests are determined from the same data and corresponding statistics and are based on the same model assumptions. A level α two-sided hypothesis test rejects the null hypothesis $H_0: p = p_0$ whenever the test statistic falls outside a level $C = 1 - \alpha$ confidence interval for p; a level α one-sided hypothesis test rejects the same null hypothesis whenever the test statistic falls outside a level $C = 1 - \alpha$ confidence interval for p.

Errors and the power of a test

- **errors in hypothesis testing**
  If *we mistakenly reject a null hypothesis that is in fact true*, we have committed a **Type I error**; in disease testing it is the error of a false positive. This will happen when we select a sample that produces an unusually low P-value; it occurs with probability exactly equal to α, the level of significance for the test.

  If *we mistakenly fail to reject a false null hypothesis*, we have committed a **Type II error**; in disease testing it is the error of a false negative. This will happen when the value of the population parameter is different from the hypothesized value, but we select a sample that produces a P-value large enough to fool us into believing the incorrect null hypothesis; we label the probability of this event β, and note that its value is very hard to determine because the true value of the parameter is unknown.

	H_0 is true	H_0 is false
reject H_0	**Type I error**	correct decision
fail to reject H_0	correct decision	**Type II error**

- **power of a test**
  probability that the test will correctly reject a false null hypothesis (see Fig. 21.4, p. 484)

  Since

  $$P(\text{fail to reject } H_0 | H_0 \text{ is false}) = P(\text{Type II error}) = \beta$$

  it follows that the power of the test is

  $$\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta.$$

  Unfortunately, this analysis does not make the power of the test any easier to determine!

- **effect size**
  the distance between the hypothesizes value of the parameter and its true value; the larger the effect size, the greater the power of the test and the smaller the Type II error

- **reducing errors**
  An investigator is given control over α, *but not over* β (recall that β depends on the unknown true value of the parameter). Decreasing α requires the evidence against the null hypothesis to be very strong in order to reject it, but this simultaneously *increases* β. In order to decrease both types of error, one must reduce the standard deviation $SD(\hat{p})$ of the sampling distribution by increasing the sample size n.