Lecture 16

Now that we've learned how to make decisions based on hypothesis testing we need to look at the possible errors in those decisions.

Types of Errors in Hypothesis Tests

In hypothesis tests, the P-value summarizes the evidence about H_o. A P-value such as 0.001 casts strong doubt on H_o being true, because if it were true the observed data would be very unusual.

When we need to decide whether the evidence is strong enough to reject H_o, we've seen that the key is whether the P-value falls below a pre-specified significance level (sometimed called α "alpha"). In practice we choose $\alpha = 0.05$ and reject H_o if the P-value ≤ 0.05. We do not reject H_o if the P-value > 0.05. The smaller α is the stronger the evidence must be to reject H_o. To avoid bias, we select α before looking at the data.

Two Types of Errors

Notice that we are basing our test only on the results from one sample. Because of sampling variability, decisions in significance tests always have some uncertainty. A decision can be in error.

Tests can have two types of potential errors:

Type I error occurs when you reject H_o when it really is true.
Type II error occurs when you do not reject H_o when it really is false.

Example:
Remember the anorexia example testing whether the therapy affected weight change. The P-value we got was 0.02 for the test of no weight change $H_o : \mu = 0$. With significance level $\alpha = 0.05$, we rejected H_o and concluded that anorexic girls have a positive mean weight change when undergoing therapy. If the therapy truly has a positive effect, this is a correct decision. But what if, in reality, the therapy has no effect, and the true population mean weight change (unknown to us) is actually 0. If actually the anorexia therapy has no effect (that is, if μ=0), we've made a Type I error. H_o was actually true but we rejected it based on our data from the one sample.

Example:
Remember the astrology example testing whether the astrologers could make correct predictions better than random. In that study $H_o : p = 1/3$ corresponded to random guessing by astrologers. we got a P-value of 0.40. with significance level = 0.05, we do not reject H_o. If truly $p = 1/3$ then this is the correct decision. However, if astrologers really can predict better than random guessing (so that $p > 1/3$), we've made a Type II Error, failing to reject H_o when it is truly false.

When we made a decision based on a hypothesis test there are always four possible results, two of which result in errors. In practice, when we plan to make a decision in a hypothesis test, it is important to know the probability of an incorrect decision.

		Decision	
		Do not reject H_o	Reject H_o
Reality about H_o	H_o true	Correct decision	Type I error
	H_o false	Type II error	Correct decision

Analogy to a legal trial:

Consider a decision in a legal trial. The null hypothesis is that the defendant is innocent. The prosecution has to find evidence to prove the the null hypothesis is false in favor of the alternative hypothesis that the defendant is guilty. The jury rejects H_o if the evidence is sufficient. A Type I error, reject a true null hypothesis, occurs when we convict a defendant who is actually innocent. Not rejecting H_o means that the defendant is acquitted (judged not guilty). A Type II error, not rejecting H_o even though it is false, occurs when we acquit a defendant who is actually guilty.

		Decision	
		Acquit	Convict
Reality about defendant	Innocent (H_o true)	Correct decision	Type I error
	Guilty (H_o false)	Type II error	Correct decision

The significance level is the probability of a Type I error

With significance level $\alpha = 0.05$, we reject H_o if the P-value ≤ 0.05. For a two-sided test, the two-tailed probability that gives us a P-value of exactly 0.05 is when our test statistic $|Z| \geq 1.96$. We call this area on the Z table the **rejection region**. The values in this region represent the values of a test statistic we'd least expect to observe if H_o were true.

Now if H_o is actually true, the sampling distribution of Z test statistic is the standard normal distribution and therefore, the probability of reject H_o (which is the probability that $|Z| \geq 1.96$) is exactly 0.05 - precisely the significance level.

Suppose H_o is true.
The probability of rejecting H_o, thereby making a Type I error, equals the significance level for the test.

So, when we make a decision why don't we use an extremely small probability of Type I error, like $\alpha = 0.000000001$? For instance, why don't we make it almost impossible to convict some who's really innocent?

Well, when we make α smaller, we need a smaller P-value to reject H_o. It then becomes harder to reject H_o even when H_o is truly false. The stronger the evidence needed to convict someone, the more likely it becomes that we will fail to convict defendants who are actually guilty. In other words, the smaller we make the probability of Type I error, the larger we make the probability of Type II error.

As P(Type I Error) goes down, P(Type II error) goes up.

Limitations of Hypothesis Tests

Tests and confidence intervals based on statistical significance are used in industry, science, courtrooms, marketing, etc. They are valued because they are capable of succinctly quantifying the chance of the results of studies and experiments. We've alluded to some of the dangers/shortcomings of using tests and CI's. In this section, we'll get very specific about four mistakes you should avoid making.

Do Not Live and Die Based on the p-value

How small a p-value is convincing enough to reject the null?

Consider these two things:
-How plausible is H_o? If H_o represents a long-held belief of many people, then you will need a very small p-value to be convincing.
-What are the consequences of rejecting H_o? Sometimes rejecting H_o will mean a lot of money will have to be spent in changing over production lines of marketing plans.

Different people and different fields can vary in what they consider to be acceptable, but in many fields today, there is a well-accepted α level in use. Just realize that there is not a sharp boundary between "significant" and "insignificant", only increasingly strong evidence against H_o as the p-value decreases.
Example:
You run your data to test a hypothesis about the mean μ. The p-value is .051, but the α-level chosen beforehand is .05. Do you not reject H_o?

Beware of Multiple Analyses

Statistical inference works best when you know what effect you are looking for, design a study to research that effect, and use a confidence interval or test to weigh the evidence.

What <u>not</u> to do:
Sometimes people will have a large data set and will run hundreds of tests on the data, looking for one that produces something statistically significant. DO NOT DO THIS!!!

Example: A researcher runs 500 hypothesis tests at the α=.01 level and finds four that are statistically significant. Can she be sure that there really is evidence to reject H_o in those four cases?

NO. By the laws of probability, 1% of the 500 tests will turn out to be significant, but they may not really be. 1% of 500 is five. The researcher found four tests were statistically significant. Pure chance could have made those particular tests turn out to be statistically significant. To determine whether there really is a difference or not, she needs to resample from those four populations, and run the tests again. If again they all turn out to be statistically significant, then she has evidence that the results were not due simply to chance.

Solution: Do exploratory data analysis and look for patterns. When you develop a hypothesis, design a study to test it alone.