STAT371: Introductory Applied Statistics for the Life Sciences **Spring 2018**

Chapter 6 — Hypothesis testing

1. Hypothesis Testing Definitions

   The first form of inferential statistics was estimation, and the goal was simply to make a good guess, or range of guesses, about a population parameter. In this chapter, we wish to make a guess about a parameter before gathering data, and then see if the data is consistent with our guess. There are many ways to approach testing, but we will present a common one, the Neyman-Pearson paradigm. We start with some defintions:

   - In hypothesis testing, there are two competing hypotheses. One hypothesis, called the **null**, usually denoted H_0, is the uninteresting result. The null is assumed true unless there is evidence to the contrary. Usually the null specifies a single value for a parameter. The other hypothesis, called the **alternative**, usually denoted H_A, is usually what we would like to show. The alternative usually specifies a range of possible values for the parameter.

   - Data are gathered, and a quantity is computed called the **test statistic**. The specific formula for the test statistic will depend on the parameter being tested and the nature of the sampling. The test statistic is an RV. The numerical value of the test statistic (its realization) will be used as evidence to decide between the null and alternative.

   - If the test statistic indicates that the null is likely false, we say that the statistic falls in the **rejection region** and we **reject** the null. If the test statistic offers insufficient evidence against the null, we say that we **do not reject** the null.

2. The types of errors that we could make when doing a test is best shown in a table:

	Reject H_0	Not Reject H_0
H_0 True	Type I Error / α	Correct
H_0 False	Correct	Type II Error / β

   If the null is true and we do not reject it, or if the null is false and we reject it, we have done the right thing. If the null is true and we reject it, we have made the wrong conclusion. We say we have made a Type I error, or an α error. We also call α the probability of a Type I error, which can be expressed as follows:

$$\alpha = P(Reject\ H_0 | H_0\ true)$$
or
$$\alpha = P(\text{Test statistic falls in the rejection region } | H_0\ true)$$

We read the "|" character in the above expressions as shorthand for,"conditional on," or "assuming that," or "given that." In order to compute this probability, we need to know the distribution of our test statistic when the null is true. Smaller values of α indicate a better test.

If the null is false and we do not reject it, we have also made the wrong conclusion. We say we have made a Type II error, or a β error. We also call β the probability of a Type II error, which can be expressed as follows:

$$\beta = P(Not\ reject\ H_0 | H_0\ false)$$
or
$$\beta = P(\text{Test statistic does not fall in the rejection region } | H_0\ false)$$

Since the alternative usually specifies a range of values, this probability can only be computed by specifying a single value of the parameter that falls within the alternative range. It is really a collection of probabilities. Thus, usually β is given a subscript that indicates the specific value of the parameter under which the probability is computed. Smaller values of β indicate a better test.

Very closely related to β is **power**:

$$\text{Power} = 1 - \beta = P(Reject\ H_0 | H_0\ false)$$
or
$$\text{Power} = 1 - \beta = P(\text{Test statistic falls in the rejection region } | H_0\ false)$$

As with β, power can only be computed given a single value of the parameter that falls in the alternative. Higher power indicates a better test.

We've said that we desire a small α and a small β (or equivalently, large power). Unfortunately, for a given fixed sample size, if we adjust our rejection region to decrease α, β will go up, and vice versa. The only way to decrease both α and β simultaneously is to increase the sample size. Thus, when deciding on a rejection region, we must carefully consider the relative importance of the two types of error.

3. We now discuss one more general topic regarding hypothesis testing.

   - The **p-value** is defined to be the probability of a test statistic realizing to a value that is as or more extreme than the one actually observed, when the null hypothesis is true. Smaller p-values indicate relatively more evidence against the null hypothesis.

   - The p-value required to cause a rejection of the null is called the **significance level** of the test.

   In most situations, reporting the p-value so that it may be used as the degree of evidence against the null is better than only stating the reject or not-reject decision. When the p-value is reported, each individual can use it to make their own decision about whether the evidence is strong enough to reject the null or not.