Hypothesis Testing

Recall that there were three basic forms of inferential statistics: estimation, testing, and fitting. We've just finished estimation. We now discuss some basic concepts of the second form, hypothesis testing.

1. Hypothesis Testing Definitions

The concepts in this section are covered in sections 5.1, 5.4, and 5.6 of Ott and Longnecker.

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 section, we wish to 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.

As a non-statistical example, consider a fire alarm. For a fire alarm, natural choices for the null and alternative are:

H_0: There is no fire.
H_A: There is a fire.

Notice that the null hypothesis is the uninteresting case, that causes no further action. Also note that these hypotheses do not overlap, and that in this case, these two hypotheses cover all possible occurrences. One choice of test statistic might be the temperature in the room. Generally, a higher temperature would be more evidence against the null. We could equally well use the concentration of smoke particles in the room as a test statistic. Again, more smoke particles would be more evidence against the null. For simplicity, let's continue by choosing temperature as our statistic.

A rejection of the null would indicate that we believe a fire is occurring and the alarm should go off. A not-rejection would indicate that we believe there is no fire. If we were actually constructing a fire alarm, we would need to decide a cutoff for temperature that determines when the alarm goes off. At what temperature does it become unlikely that conditions are typical when there is no fire? In this example, perhaps that would come from a study on room temperatures under typical conditions. Statistically, this would be equivalent to knowing the probability distribution of the test statistic conditional on the null hypothesis being true.

Call our test statistic X. Suppose research indicates that a temperature over 110F indicates the possibility of a fire. Then our rejection region would be $X > 110$. Then, if we measured the temp in a room and we got, for example. $x = 70$, we would conclude that things were fine, and would not reject the null. But, say we got $x = 200$. Then we would reject the null and the alarm would go off.

This brings us to discussion of the types of errors that we could make when doing a test. This 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. In the fire alarm example, α is the probability that the fire alarm goes off when there is no fire, or said differently, the probability of a false alarm.

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. In the fire alarm example, β is the probability that the alarm does not trigger when a fire is actually happening.

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. In the fire alarm example, power is the probability that the alarm correctly goes off when a fire is happening.

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. For the fire alarm example, the choice is quite obvious. It is much worse to have the alarm not go off when there is a fire, because this could result in serious injury or death. False alarms are annoying, but are not dangerous. So it is much more important to choose a rejection region that gives us a reasonably large power. In other situations, a different choice may be appropriate.

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.

In the next section, we will give examples of some specific tests based on samples from one population. This will allow us to see these concepts in action.