

**MthStat 465, Spring 2005, Lecture Number 20**  
**Overview of Testing Statistical Hypotheses, Part I**

We want to establish the terminology for the formal process of testing statistical hypotheses. We will illustrate the concepts with coin tossing.

Suppose we are given a sample space  $S$ , a set of events,  $\Sigma$ , for this sample space, and a collection of possible probability measures,  $\{\Pr_a, a \in A\}$ . (Here  $A$  is an indexing set, such as the real numbers, or the integers, etc.) The sample space and set of events are intended to model some phenomenon. For example,

$$S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$$

and  $\Sigma =$  all  $2^8$  subsets of  $S$ . There is an uncountable number of different probability measures we could define for this set of events.

Back to our general discussion. Among all the probability measures, we single out some, which we call the **null hypothesis**. Which probability measures we designate as the null hypothesis depends on the nature of the phenomenon we wish to model and our beliefs about this phenomenon. Having decided on the null hypothesis, we would perform the experiment, get a sample point, and see if this sample point is consistent with the null hypothesis. In many cases the null hypothesis consists in only one probability measure, but this is neither always the case nor always desirable. The null hypothesis is usually denoted by  $H_0$  (H for hypothesis and 0 for null.)

In our coin tossing model we might believe that the coin is fair and that the tosses are independent. If so, we would take the null hypothesis to be the probability measure that assigns to each single element event the probability  $1/8$ . We would then toss the coin in question 3 times, record the outcome, and see if this is consistent with the null hypothesis. For example, the sample point might be  $(HHH)$ . We have to decide if this is consistent with the null hypothesis. When the null hypothesis consists of only one probability measure it is called a **simple null hypothesis** and when it contains more than one probability measure it is called a **compound null hypothesis**.

The goal of hypothesis testing is to make precise this idea of the sample point being consistent with the null hypothesis. In order to do this, we have to know what the other possibilities are, because we want to be able to say "consistent compared with ...". In mathematical language, we have to designate some (possibly all) of the other possible probability measures as alternative explanations. This collection of other probability measures is called the **alternative hypothesis**. When the alternative hypothesis consists of just one probability measure, it is called a **simple alternative hypothesis**, and when it consists of more than one probability measure it is called a **compound alternative hypothesis**. There are special types of compound alternative hypotheses that have more structure to them, as we will see in later lectures. The set of alternative hypotheses is usually denoted by  $H_1$ .

In the coin tossing example, the alternative hypothesis could consist of one or many other probability measures. For example, if we let  $\Pr_p$  denote the probability measure on  $S$  that assigns probability  $p^x(1-p)^{3-x}$  to the one element set containing an outcome with  $x$  H's and  $3-x$  T's, then we might have  $H_1 = \{\Pr_p : p \in (0, 1/2) \cup (1/2, 1)\}$ , or  $H_1 = \{\Pr_p : p = 1/3 \text{ or } p = 4/7\}$ . The important point is that  $H_1$  can contain any probability measure except the ones in the null hypothesis.

After we have run the experiment and collected our sample point, we make one of two judgements:

- The sample point supports the null hypothesis;
- The sample point does not support the null hypothesis.

There are of course two possibilities in reality:

- The null hypothesis contains the right probability measure for the experiment;
- The null hypothesis does not contain the right probability measure for the experiment;

There are two obvious errors on our part:

**Type I error:** We declare that the sample point does not support the null hypothesis when, in fact, the null hypothesis contains the probability measure that describes the experiment;

**Type II error:** We declare that the sample point supports the null hypothesis when, in fact, the null hypothesis does not contain the probability measure that describes the experiment.

As a practical matter we would like to avoid making these errors. Since they cannot be avoided entirely, we try for the next best thing, which is to make them unlikely, that is, we want to choose a decision rule that does two things:

1. Make the probability of Type I error small under the assumption that the null hypothesis contains the probability measure that does describe the experiment. This probability is called the **probability of Type I error**.
2. Make the probability of Type II error small under the assumption that the alternative hypothesis contains the probability measure that does describe the experiment. This probability is called the **probability of Type II error**. 1– the probability of Type II error is called the **power** of the test against the alternative.

In many cases we can only control the probability of Type I error, and the power of the test is simply a function of the elements of the alternative hypothesis. Our goal, then, is to choose our decision rule by requiring that the probability of Type I error does not exceed a given threshold. This threshold is called the **level** of the test. Typical choices of level are 0.10, 0.05 and 0.01. These are chosen more by tradition than anything else.