Testing Statistical Hypotheses

The Pennies experiment is intended to introduce you to the ideas involved in statistical hypotheses testing. We begin with a question concerning the fairness of the coins, perform an experiment to gather data on that question, and then need to arrive at some inference based on those data. Below I will lead you through some thoughts and terms connected to the process.

1. **The Hypotheses**
   The nature of the question to be addressed in a statistical test is of paramount importance. There is a tremendous difference between an interest in asking "Is the coin fair?" or asking "Is the coin biased?" Statistical techniques can be used to confirm (in a statistical sense) that a particular coin is biased, but it is not possible to prove that a coin is fair, if by fair it is meant that the coin will land heads with a probability exactly equal to 0.5. We will address the implications of this later in the semester. But for now, realize that distinguishing between probability of heads equal to $p_1 = 0.4999999999999$ and $p_2 = 0.50$ would be extremely difficult. Yet a coin with probability of heads p_1 is in fact a biased coin, as would be any coin without probability of heads exactly 0.5.

   While coins which don't have probability of heads exactly 0.5 are indeed biased in the strict sense, some small deviation from 0.5 might well be acceptable from a practical point of view in each particular application. But this is a digression that we will address later. For the present we will restrict attention to testing for differences, not showing equalities. In this particular problem then, our interest is in showing that the coin is biased, not that it is unbiased.

   We begin our testing procedure by defining what is known as the null hypothesis; that result which we hope to be able to show is false from our experiment. In the coin experiment our null hypothesis is that the coin is unbiased. That is, we hope to be able to show the coin is biased, so we make our null hypothesis that the coin is unbiased.

   If the null hypothesis is what we hope to show as false, then the options left over are what we hope to be able to show as true. This is the alternative hypothesis. Hence the alternative hypothesis is what we hope our data will be able to confirm.

2. **The Actions**
   There will be two actions which the statistician can take in a testing of hypotheses problem. One action is to reject the null hypothesis. This action is taken when the statistician feels that the data support the truth of the alternative hypothesis, that is support the fact that the null hypothesis is not true. This is the action which was desired for the final results of the experiment.

   It is natural to think if one doesn't reject the null hypothesis, then one must accept it as true. Since the statistician is limited to two actions, people often refer to this second action as accepting the null hypothesis. However, this is improper terminology.

There are two alternatives to rejecting the null hypothesis. One can accept the null hypothesis as true, or not have enough information to feel comfortable claiming it to be either true or false. It is this area of not being able to make a decision which is often forgotten, and leads to improper interpretations to the action often referred to as \accepting the null hypothesis." For this reason I prefer to refer to this latter action as **not rejecting the null hypothesis** which allows for the inclusion of no decision.

3. **The Mistakes** In an application of testing hypotheses one of two states of nature will prevail. Either the null hypothesis will be true, or the alternative hypothesis will be true. In our example, either the coin will be unbiased, or it will be biased. Unfortunately, we don't know which case is the true state of nature.

   Based on our data and our strategy of what to do with these data we will take one of two actions, either reject the null hypothesis, or not reject it. If we reject, and in fact the alternative is true, we have taken the proper action. But if we reject the null, and in fact it is true, then we have committed a type I error. Since rejection of the null is a positive act in that we are willing to say that we believe the alternative to be true, we pay close attention to the probability with which we are willing to make a type I error.

   In application of statistical procedures, we pick a technique or tool for which the probability of a type I error can be computed. What this means is that we can set rules for rejection of the null hypothesis which would have us incorrectly rejecting the null with as small a probability as we choose. Of course, once the data has been gathered and the action taken, either we have made a mistake or we have not. But this probability with which we are willing to incorrectly reject the null is often referred to as the \ level of signi¯cance " of the test. Another term connected with the type one error is called the p value. This is the probability that we would observe data as or more extreme in support of the alternative, if in fact the null hypothesis were true. We will talk about these terms in relation to the pennies data.

   The second type of error that one can make in testing is to not reject the null hypothesis when in fact the alternative is true. This type of error is frequently ignored by experimenters, probably because it is not well understood. The probability with which we can expect to reject the null hypothesis is called the power of the test procedure. You experienced this with the pennies in that you were not able to judge that there was a bias after only three coins, but by the time you were at ¯fty it was clear that the bias was there. The larger number of trials provided more power to statistical test.