Testing of Hypothesis

(The following section is largely based on the relevant section of Lavine's book.)

An important reason for conducting experiments is to test hypothesis. Typically, the purpose is to see if there is enough evidence to reject a hypothesis that says that "nothing interesting is happening". This is called the null hpyothesis. In addition an alternative hypothesis is formulated in order to make clear what one means by "something interesting is happening". As we shall see the *formulation* of the alternative hypothesis allows us to clearly see a dividing line. Thus, the two hypothesis should be exclusive, but *need not* be exhaustive. Some examples will clarify the issue. (We use the standard convetion that H_0 denotes the null hypothesis and H_a denotes the alternate hypothesis.)

- Medecine
  - H_0: the new and old drug are equally effective.
  - H_a: the new drug is more effective
- Physics
  - H_0: Newtonian mechanics holds
  - H_a: Quantum mechanics holds
- IISER
  - H_0: IISER education makes no change to students
  - H_a: IISER education adds value to students
- Coaching
  - H_0: Coaching classes have no effect on JEE performance
  - H_a: Coaching classes improve JEE performance

In order to use statistics to carry out hypothesis testing we must:

- Formulate H_0 and H_a and think of an experiment that will distinguish the two. This will give a sequence of (independent, identical) random variables when we repeatedly carry out the experiment.
- Formulate a statistic $w(X_1, \ldots, X_n)$ (recall that a "statistic" is a function of the data) such that under H_0 the expected value of this statistic can be computed or simulated. Moreover, the expected value of this statistic under H_a will be different.
- Calculate the distribution of w under the assumption H_0.
- Check if the observed distribution of w is sufficiently close to the computed one.

In order to understand the steps, let us work on the JEE example. Note that we will do the most elementary type of statistics. We will ignore issues of confounding factors/variables, how random sampling is done etc.

Let us take the scores of students who have taken coaching as a sequence of independent identical variables $X_1, X_2, \ldots, X_n$ distributed as per some normal distribution $N(\mu_1, \sigma_1^2)$.

Let us take the scores of students who have *not* taken coaching as a sequence of independent identical variables $Y_1, Y_2, \ldots, Y_n$ distributed as per some normal distribution $N(\mu_2, \sigma_2^2)$.

We expect that $\mu_1 \neq \mu_2$. Hence, we take $w = \overline{X} - \overline{Y}$ as the difference of the average scores. (In other words, $\overline{X} = (\sum_i X_i)/n$ and $\overline{Y} = (\sum_i Y_i)/n$.) Under H_0 we expect $w = 0$, while under H_a we expect $w > 0$.

Under the hypothesis H_0, if n is large enough, by the Central Limit Theorem, we can approximate the distribution of w by $N(0, \sigma_w^2)$. Here, the theoretical value of σ_w^2 can be calculated by using our assumption that the X's and Y's are independent, we see that

$$\sigma_w^2 = \sum_i \sigma^2(X_i/n) + \sum_i \sigma^2(-Y_i/n) = (1/n^2)\left(\sum_i \sigma_1^2 + \sum_i \sigma_2^2\right) = (s_1^2 + s_2^2)/n$$

Now, after carrying out the experiment, we can use the sample variance s_1^2 of X_i's (respectively s_2^2 of Y_i's) as a reasonable value for σ_1^2 (respectively σ_2^2). Thus, we can calculate $s^2 = (s_1^2 + s_2^2)/n$ and use this value of s to get an interval $[-2s, 2s]$ (or $[-3s, 3s]$) within which (assuming H_0) we expect w to lie. We calculate w as the difference of the sample means and check whether this is true.

If it is true, then we *cannot reject* the null hypothesis. In other words, we have not found adequate evidence that coaching classes affect the JEE score.

On the other hand, if w is *not* in the interval $[-2s, 2s]$ we can say that we do have significant evidence that coaching classes do affect the JEE score.

Note that we *cannot* assert that H_a is true, or even that we have evidence for H_a. In fact, H_a was used at only one point in the above formulation: in order to make sure that w did indeed give a different value under H_a than under H_0.

Another method to test the same hypothesis is as follows. We assume that we can test each student *before* they enter coaching as well as afterwards. (For example, we can conduct a mock JEE test.) Let X_i denote the score of the i-th student before coaching and Y_i denote the score of the same student after coaching. We now define a random variable Z_i to be 1 if $Y_i > X_i$ and 0 otherwise. We think of Z_i as independent identical Bernoulli random variables with probability of success as p.

Let w be the statistic $\overline{Z}$ which counts the number of "successes". Under the null hypothesis H_0 we have $p = 1/2$ since there would be a random fluctuation of the score in that case. On the other hand, under H_a we think $p > 1/2$. Thus, we can use the Binomial distribution for w, or for large n we can approximate it by $N(n/2, n/4)$.

Specifically, if $n = 100$, we note that $\mu = n/2 = 50$ and $\sigma^2 = n/4 = 25$, so $[\mu - 2\sigma, \mu + 2\sigma] = [40, 60]$. Thus, if there are between 40 and 60 students who do better in the second test, then we have *no evidence* that coaching classes do anything for the students performance in JEE. On the other hand, if more (or

less!) students do better in the second test, then we do have evidence that the coaching is having an effect.

It is tempting to say that if there are more students doing well, then there is evidence to support a positive effect of coaching, but note that this is is not how the current test has been formulated. So such a conclusion needs to be supported by a test designed to test *this* hypothesis.

Errors in Testing

Errors in testing are often classified into two broad types:

Type I error Rejection of the null hypothesis when in fact the null hypothesis is true.

Type II error Failing to reject the null hypothesis when in fact the null hypothesis is false.

It is good to be aware of reasons for such errors.

1. Not being aware of the sample size. If the experimenter does not state the size of the sample, then errors are possible. If a person examining the results presented does not ask for the sample size, then that person is not being critical enough.

2. Ideally, the sample size should be decided in advance based on the assumptions about the nature of the experiment. In many cases, one may not be able to acquire a sample of the requisite size easily. In that case the results should be called preliminary.

3. Interpreting a double negative as a positive. Failure to reject the null hypothesis does not imply that the alternate hypothesis is verified or even probable. One should devise a probability distribution associated with the new hypothesis, set *it* as the null hypothesis and obtain a positive result about it.

4. Post facto calculations of confidence levels. For example, if the deviation is greater than $3s$ then this does not justify the statement that the confidence is greater than 99%! The value of s is also the result of the experiment and is therefore the value obtained from a random variable.

5. Ignoring prior probabilities. It may be tempting to assume that all values of some parameter are equally likely. However, earlier experimental results may point to varying probabilities.

6. Ignoring Likelihood ratios. As seen earlier, when likelihood ratios are of the order of 10-20, there is little to chooce between two possible values of the parameter.