Testing of Hypotheses

One of the most important and yet one of the most controversial areas of application of statistics is to the testing of hypotheses.

One common method of science requires that the theorist comes up with propositions that can be "falsified" by carrying out experiments. Such a proposition is (in this context) called a hypothesis.

We must now *design* an experiment to *test* (actually falsify!) the hypothesis.

In many cases, the experimental process is subject to stochastic error, so the result of the experiment will be a random variable. Hence, we must carry out the experiment a large number of times with the idea of applying the law of large numbers or the Central Limit Theorem in order to increase the probability of getting a better decision (on the falsification or otherwise of the hypothesis).

In many situations, the theoretical framework asserts that there is a "pattern" in nature. Thus the falsification of this is the assertion that there is *no* patterm to the results. In other words, this is the assertion that the experimental results are "random". This is often interpreted as saying that the experimental results are a random variable drawn from some standard distribution like the uniform distribution (discrete or continuous), Poisson or Normal distribution. This opposition of the theory is called the "null hypothesis" which is to be contrasted with the "alternate hypothesis" that something "interesting" is taking place.

We analyse the distribution based on the null hypothesis and decide on a size of the sample (the number of repetitions of the experiment) and a consequent a-confidence interval (for a value of a which we need like 0.68, 0.95 or 0.99). We then carry out the experiment and compute the *sample* mean and *sample* variance as estimations of the confidence interval. If the *theoretical* mean (the value from the null hypothesis) lies in this interval, then with confidence level a, we *cannot* reject the null hypothesis. However, if it does *not* lie in this interval, then with confidence level a, we can assert that the *null* hypothesis can be rejected.

Some examples will help to clarify the above framework.

Choice of Majors

The BS-MS students are asked to pick a major once they finish their core courses. They pick the major based on their subject of interest. Hence, we may feel that they should do better in that subject as compared with the remaining subjects.

The null hypothesis is that students perform equally well in all subjects.

We can only make these measurements for all students who have graduated from IISER Mohali so far. This sample size is 228. Ideally, we should be able to pick

our sample size. We will see below that this is one of the sources of errors in statistical hypothesis testing.

We measure the difference between the performance index in major specific courses versus the rest of the courses. The null hypothesis means that this difference is 0. The actual value is $m = 0.16$!

Now every student does a little better than their average score (CPI) in some courses and worse in some courses. The *root mean square* deviation between the students cpi and the performance in a particular course, taken over all courses and all students, is a good measure of the standard deviation of the performance index; this turns out to be $\sigma = 1.81$. We divide by the square root of the sample size to get an approximate standard deviation of the quantity we are measuring; this gives $s = 0.12$.

Now the 99% confidence interval $[m - 3s, m + 3s]$ turns out to be $[-0.20, -0.52]$ which does contain the value 0.

In other words, we fail to reject the null hypothesis that the students do about the same in all courses as they do in their chosen subject!

We may be tempted to say that they do *better* in their chosen subject based on the fact that the mean is positive! This temptation to redefine or re-interpret the experiment after it is performed is another source of statistical errors which we will see later.

In this particular case, the positive value of the mean may be due to the possibility that those who performed better performed significantly better! In any case, we need to set up a different theoretical model before making any conclusions.

Another way of looking at this is that our null hypothesis only says that there is *no* difference between the performance in the chosen subject at other subjects. Hence, this hypothesis is perhaps incorrectly formulated. Poorly chosen null hypotheses are one of the major reasons for erroneous conclusions of statistical testing.

Did it Work?

After the second mid-semester examination, we had a revision of the notion of convergence and then a quiz on it. Let us compare the performance in this quiz with the performance in question 3 of the second mid-semester examination.

We take an (anonymized) list of students along with their marks in question 3 in the second mid-semester examination (denoted `sq3`) and there marks in quiz number 7 (denoted `qq7`). (Note that unavailable entries are denoted by `NA`.)

```
> adat[1,]
fq1 fq2 fq3 fq4 fst sq1 sq2 sq3 sq4 snd qq1 qq2 qq3 qq4 qq5 qq6 qq7 qq8
  1   2 3.5   2 8.5 NA  NA  NA  NA  NA    1   5   1   0 0.5   2  NA   2
```

We now prepare the list that verifies the truth or falsity of our hypothesis.

```
> hyp <- adat$qq7 >= adat$sq3
> hyp <- hyp[!is.na(hyp)]
```

The second line is useful as a way of discarding the `NA` cases.

If the extra revision session on "Convergence" had no effect (our null hypothesis), then the value of `hyp` would equally likely be true or false. Hence, we wish to compare `hyp` with a sequence of coin flips.

```
> sum(hyp)
119
> length(hyp)
155
```

Now we can either apply the solution of the assignment to calculate the divergence from a sequence of coin flips or we can calculate the confidence interval using `R`. We do the latter here:

```
> m <- mean(hyp);m
[1] 0.7677419
> s <- sd(hyp)/sqrt(length(hyp));s
[1] 0.03402771
> c(m-3*s,m+3*s)
[1] 0.6656588 0.8698251
```

Hence, we see that 0.5 is not in the 99% confidence interval. Thus, we can assert with 99% confidence that the extra lecture on convergence had an effect on the class. We can even assert that the effect is positive!

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.