Multiple Comparison : A Simple Example

Samahriti Mukherjee
Roll No.: BS2003
B.Stat 3rd year
Indian Statistical Institute

1 Introduction

Why are we worried about multiple hypotheses testing in context of a linear model? Let us take a very simple example, where we have a single factor input and we are interested in testing a very standard hypothesis.

2 Model

Our model is of the form

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad\qquad \vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 I)$$

where $\sigma^2 > 0$ is unknown.
We are interested in knowing whether all the α_i's are same, i.e., our null hypothesis is :

$$H_0 : \alpha_1 = ... = \alpha_I$$

Suppose we have a setup where we are trying to compare different varieties of the same crop. So your α_i measures the effect of the ith crop and in this case we are interested whether the different varieties at all differ in terms of their yields or not. So if they do not differ at all in that case this null hypothesis will be true, that all of the varieties are basically the same, which is the uninteresting situation. If you find that the different varieties are really producing different amount of yields in that case that will be the discovery of an interesting thing. Possibly more time and money will be required in cultivating one particular variety than another. So even if at least one pair produces different amount of yield, you are going to reject the null hypothesis. If all the varieties are producing different yields, that is a different place. So basically, we are interested in comparing each possible pair. So here we are splitting up the same H_0 in $\binom{p}{2}$ ways. So if we say that i takes p values then there are p different varieties that we are comparing. In that case we have got $\binom{p}{2}$ many null hypotheses that we are testing. These are components of null hypothesis that together constitute the big null hypothesis that we have where we compare $\alpha_s = \alpha_t$ for each pair (s, t), $s < t$.

These are not all independent. (e.g. $\alpha_1 = \alpha_2$, $\alpha_2 = \alpha_3 \implies \alpha_1 = \alpha_3$). These are the conclusions that we want to draw. So we want to compare each variety with each other variety. So instead of coming up with a single accept or reject we want to say for each such comparison whether we are accepting it rejecting it. So our null hypothesis will be:

$$H_0^{(s,t)} : \alpha_s = \alpha_t, \qquad k = \binom{p}{2}$$

where k is the number of hypotheses we are testing.
That is a more useful output in practice. That is why we are interested in multiple hypothesis testing in the setup of Linear Model.

3 A Simple Example

We know that testing multiple hypotheses is not same as testing k hypotheses in parallel. Let us take a very simple example to appreciate that. Suppose we have got $k = 100$ coins and we are interested in testing

$$H_0^{(i)} : i\text{-th coin unbiased}$$

So our overall hypothesis is all the coins are unbiased. So we split the hypothesis into k many hypotheses. The alternative will be that the coin is biased. We must have some data. So each coin is tossed 50 times, independently.

We want FWER$\leq$5%, so we impose family-wise criteria, not separately on individual tests.

Now let us run separate tests for individual hypothesis. Suppose we carry out each of these tests at a 5% level, so for each one we get acceptance/rejection. Will that work? No.

When we say that we are going to work with each coin separately and carry out the test at 5% level, we mean that even if the coin is unbiased, we are going to reject it 5% of the time. Now we have 100 coins. So we are going to reject around 5 of those coins, so we will say that 5 coins are actually biased. Now, what does this say about our **False Discovery**? On an average we make 5 False Discovery, and **FWER** says even if you make at least one False Discovery that is bad. So you are going to make at least one false discovery with a very high chance, possibly you will have something like 99% chance that you will make a False Discovery, because if the average is 5, then the random variable taking the value 0 is pretty low. So you will end up having something whose FWER is much much higher than 5%. So this shows why carrying out each of the individual hypothesis at 5% level is not going to give you a good answer.